DOCUMENT RESUME

ED 155 187 95 TM 007 030

AUTHOR Godbout, Robert C.; And Others
TITLE The Problem of Spurious Significance in Classroom Education Research.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Aug 77
CONTRACT NIE-C-74-0088
NOTE 171p.

EDRS PRICE MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS *Classroom Research; Educational Experiments; Educational Research; *Hypothesis Testing; Mathematical Models; Probability; *Research Problems; *Simulation; *Statistical Analysis; *Tests of Significance
IDENTIFIERS Chance Expectation Test; Nonchance Expectation Test; Statistical Power; Type I Error

ABSTRACT

The problem of spurious significance in multivariate exploratory research is discussed. When a very large number of statistical tests are performed, many tests will be significant on the basis of chance alone. To counter this problem, the use of two sign tests to analyze sets of results has been suggested; the chance expectation [CE test] assesses whether the number of obtained significant results exceeds the average number of significant results expected on the basis of chance, and the nonchance expectation [NCE test] assesses whether the number of significant results is highly improbable. The characteristics (Type I error rate and statistical power) of these two tests were studied using computer simulation procedures. Application of the CE and NCE tests was found to be reasonable for large sets of results, regardless of the dependence among results: On the basis of the simulation results, specific recommendations for application of the two tests have been proposed. If these tests identify a significant set of results, then they can be applied to smaller and smaller subsets of the results in order to identify the most likely sources of valid relationships. Example applications of the tests to five educational research investigations are given. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED155187

TM007 030

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Problem of Spurious Significance in Classroom Education Research

Robert C. Godbout, Paul T. Marston,
Gary D. Borich and Curt Vaughan

The University of Texas at Austin

The Evaluation of Teaching Program
The Research and Development Center
for Teacher Education

August 1977

The research reported herein was supported in part by the National Institute of Education Contract NIE-C-74-0088, Research and Development Center for Teacher Education. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that office should be inferred.

ABSTRACT

This report has considered the problem of spurious significance in multivariate exploratory research. When a very large number of statistical tests are performed then many tests will be significant on the basis of chance alone. To counter this problem of spurious significance, the use of two sign tests to analyze sets of results has been suggested. One of these sign tests (the chance expectation or CE test) assesses if the number of obtained significant results in a set exceeds the average number of significant results expected on the basis of chance. The second sign test (the nonchance expectation or NCE test) assesses if the number of significant results is highly improbable (i.e., probability less than α) on the basis of chance. The characteristics (Type I error rate and statistical power) of these two tests were studied using computer simulation procedures. Of special interest were the Type I error rate and statistical power of the tests as a function of the dependence among the results in a set. Application of the CE and NCE tests was found to be reasonable for large enough sets of results regardless of the dependence among the results. On the basis of the simulation results, specific recommendations for application of the two tests have been proposed.

Arguments have been made that the CE and NCE tests are applicable to sets of results from virtually any inferential statistical methods and to sets of replicated results. If these tests identify a significant set of results, then they can be applied to smaller and smaller subsets of the results in order to identify the most likely sources of valid relationships. Example applications of the tests to five educational research investigations suggest that the large majority of "significant" findings in that research may be spurious.

The Problem of Spurious Significance in Classroom Education Research

Robert C. Godbout, Paul T. Marston, Gary D. Borich, and Curt Vaughan

The relevant variables associated with successful classroom education have not been clearly identified. Investigations of classroom education have produced weak and conflicting results as to the antecedents of pupil cognitive growth (see Rosenshine, 1971) and affective growth (see Kash, Borich, & Fenton, 1976). Testament to the general confusion concerning important variables or dimensions is given by the scores of systems for observing classroom processes (see Simon & Boyer, 1970). It would appear that almost every researcher has developed a new observation system for his own use and such behavior is not inappropriate given the lack of compelling results with previous systems.

Given such a situation, it is not surprising to find a number of recent classroom education studies which have an exploratory (many variable) flavor, e.g., Borich, Godbout, Peck, Kash, & Poynor (1974); Brophy & Evertson (1974a); Calkins, Godbout, Poynor & Kugle (1976); Good & Grouws (1975); Soar & Soar (1973). To increase the chance of tapping predictive dimensions, most of these recent investigations have incorporated a number of different observation systems. Incorporating a wide range of possible predictors of educational success is a natural reaction to ignorance concerning relevant variables.

A clear implication of previous classroom research is that conceptions of the classroom education process have been grossly oversimplified. For example, the processes that control growth probably vary in complex fashion with context, i.e., grade level, subject matter, social climate. It is unreasonable to suppose that particular teaching behaviors will produce the same results in first-grade arithmetic classes as in junior high social studies classes. Certainly, other important but unidentified context variables determine the antecedents of

growth. Given such complexity, it would be ill-advised to investigate the influence of one type of variable in one study but other types of variables in other studies. If such a set of studies differ with regard to any important context variables (whether these variables are at present known to be important or not), then pooling the results of the studies will only lead to the type of confusion that past research has generated. A better plan might be to attempt to cover the gamut of possible types of predictors of educational success in each study, i.e., to examine the range of possible predictors in each context. Agreement of results across such studies may indicate general antecedants to growth while disagreement will indicate the need to assume that different antecedents are operative in the different contexts.

The large scale multivariable studies being conducted on classroom education would seem to invite multivariate statistical techniques, e.g., canonical correlation, discriminant analysis, or multivariate analysis of variance. Familiarity with multivariate techniques is rapidly increasing due to the increased availability of computer programs and practical discussions of the techniques (e.g., Bock, 1975; Cooley & Lohnes, 1971; Kerlinger & Pedhazur, 1973). Multivariate techniques provide overall significance tests for predictor-criterion relationships in general. Such overall tests have the advantage that only one or a few significance tests must be examined to determine if significant prediction of criterion performance has been achieved. Without such overall tests it would be necessary to examine, for example, the separate results for each predictor-criterion pair. Not only is this a monumental task (thousands of predictor-criterion pairings may exist) but it also may be quite difficult to determine whether all of the obtained significant results are merely a reflection of chance, i.e., whether significant (nonchance) prediction of criterion performance has been achieved. An additional advantage of multivariate

analysis is the possible identification of underlying dimensions which may each subsume several predictor or criterion variables. The resultant reduction in the number of pertinent dimensions is always welcome.

Given the advantages of overall multivariate tests, one might anticipate that large-scale studies of classroom education (e.g., Borich et al., 1974; Brophy & Evertson, 1974a; Calkins et al.,1976; Good & Grouws, 1975; Soar & Soar, 1973) would have used such methods. However, none of these investigations employed overall multivariate tests, such techniques being precluded by small sample sizes. Multivariate techniques which involve overall significance tests cannot be employed when the number of variables to be studies exceeds the sample size. This is because with more variables than subjects, no degrees of freedom are available for an overall statistical test on the entire variable set.

Large-scale studies of classroom education have been characterized by a large number of variables but small number of subjects. For example, Brophy and Evertson studied hundred of predictors, several criteria and several contexts with regard to fewer than 40 subjects (classrooms). Unfortunately the discrepancy between number of variables and number of subjects is not easily resolved. Decreasing the number of variables is difficult given ignorance of which types of variables may be relevant to classroom learning processes. Increasing sample size is prohibited by cost. There is general agreement that the appropriate unit of analysis in most classroom research is the classroom rather than the individual pupil.[1] Thus, sample sizes appropriate to overall multivariate analysis would involve numbers of classrooms exceeding the numbers of variables. Achieving such sample sizes would often involve a ten-fold or twenty-fold

---

[1]Discussions concerning the appropriate unit of analysis in classroom research can be found in Rosenshine (1971), Poynor (1974), and Glass & Stanley (1970).

increase in the number of classrooms studied. The resultant ten-fold or twenty-fold increase in the already sizable cost of a classroom education study is impractical.

Lack of knowledge and prohibitive costs dictate the many variable, small sample nature of current classroom education studies. These studies are best thought of as exploratory case studies rather than formal statistical investigations. They involve extensive study of small samples. Case studies represent the logical initial step in the scientific study of classroom education. If the results of case studies are clear and suggest a limited number of relevant variables, more formal statistical studies of these relevant variables would then be appropriate.

Analysis of case studies. Consider an example case study involving 100 predictor variables and 10 criteria. The typical analysis strategy is to calculate all 1,000 possible predictor-criterion correlations. In this case, a significant correlation does not necessarily imply the importance of the predictor variable in question. Individual results must be considered in the context of the full set of results. Say that the investigator in this example case study has chosen a significance level of .10.[2] Then an average of 100 of the 1,000 correlations are expected to be significant on the basis of chance alone. If a total of only 90 correlations is significant, then it is quite probable that any individual significant result is due to chance. In this case, a "significant" result cannot be interpreted as indicating a true (nonchance) relationship. On the other hand, suppose that 500 of the 1,000 correlations are significant. Then it is much more appropriate to interpret a "significant"

[2]It is not unusual for case studies (e.g., Brophy & Evertson, 1974b; Good & Grouws, 1975) to employ a .10 significance level rather than the conventional .05. The statistical power of case studies is limited by small sample size and a less stringent significance level may be chosen to raise power to a more satisfactory level.

correlation as being indicative of a true relationship.

The appropriate interpretation of a result significant at the α level depends upon the intent of the investigator. Suppose the investigator is interested in testing an a priori hypothesis concerning the linear relationship between two specific variables. In this case, a single result (the bivariate correlation) provides all the evidence relevant to the hypothesis of interest. A significant correlation can be interpreted as indicating a true relationship, with the confidence in the accuracy of this interpretation being $1 - \alpha$. In exploratory research, the main intention of the investigator is not testing a priori hypotheses concerning specific variables. Rather the intention of the investigator is more accurately thought of as "data-snooping," with the initial hypothesis of interest being whether nonchance prediction of criterion performance has been generally achieved. In an exploratory case study, the evidence relevant to this general hypothesis is a set of results, not a single result, and the significance of a particular result cannot be interpreted without consideration of the other results. There is a clear need for a screening technique involving an overall evaluation of the entire set of results. If the entire set of results could easily have been produced by chance, then no importance should be accorded to a predictor variable with significant bivariate results. Significant results for individual variables should be considered only when the entire set of results cannot easily be attributed to chance.

The need for overall evaluation of the entire set of variables in an exploratory study has been generally recognized. In studies allowing the use of multivariate tests, it is recommended (Cooley & Lohnes, 1971; Bock, 1975; Kerlinger & Pedhazur, 1973) that an overall multivariate test be performed for the entire variable set. Only if this test is significant should the results for individual variables or subsets of the variables be considered. Similar

screening tests are appropriate for analysis of variance as applied to factorial designs, trend analysis, and multiple comparisons. For exploratory investigations utilizing factorial designs, an overall test of between-groups differences should be performed initially. Only if this test is significant should the separate results for main effects and interactions be considered. In the case of trend analysis or multiple paired-comparisons in exploratory investigations, results for specific trends (linear, quadratic, etc.) or paired-comparisons should be considered only if a significant result is obtained for an overall test of the differences across levels of the variable.

While a similar overall screening test is clearly needed for exploratory case studies, conventional statistical techniques cannot be applied in overall fashion in such studies. The present paper was undertaken to present (1) alternative schemes for testing overall significance in case studies, (2) computer simulation tests of the reasonableness of these schemes, and (3) applications to case study investigations of classroom education.

Analysis of Sets of Results

An overall multivariate test cannot be applied to the entire data set of a case study. Thus, the hypothesis that there is significant prediction of criterion performance cannot be tested by appeal to a single overall result. Rather a set of results (all possible predictor-criterion correlations) must be analyzed collectively to determine if such a set of outcomes can reasonably be expected on the basis of chance. The common sign test offers a possible approach to such analysis of a set of results.

An excellent discussion of sign tests and their wide range of application is provided by Bradley (1968; 1969). The sign test, one of the simplest inferential statistical tests is appropriate for analyzing the set of outcomes

from N trials of an event. Each trial must have one of two mutually exclusive outcomes--State 1 or State 2. The sign test involves simple comparison of the number of obtained State 1 trials with an expected value. In applying a sign test to a set of predictor-criterion correlations, each correlation represents a trial and the two mutually exclusive outcomes of a trial are significance at the α level and nonsignificance at the α level. The sign test involves comparison of the obtained number of significant correlations in a set with an expected number of significant results for that set. The key to applying a sign test to a set of results is arriving at an appropriate expectation for the number of significant results. Two different expectations--the chance expectation and the nonchance expectation--will be considered in turn.

Chance expectation. The chance expectation (CE) for a set of results refers to a number of significant results which is likely to occur on the basis of chance. For the present purposes, the chance expectation is formally defined as the average number of significant results expected on the basis of chance. As pointed out by Miller (1966), the chance expectation for any set of results is always the number (N) of results times the significance level (α), i.e.,

$$CE = N\alpha.$$

The chance expectation test for a set of results is a sign test involving simple comparison of the actual number of observed significant results with the chance expectation. If the observed number does not exceed the average number expected by chance, then there is little reason to suspect that true relationships exist. When true relationships do exist, the observed number of significant results should tend to exceed the chance expectation. Thus, the following decisions can be made on the basis of the chance expectation test.

(1) If the observed number of significant results is less than or equal to the chance expectation, then the set of results is assumed to be chance produced.

(2) If the observed number is greater than the chance expectation, then the set of results is taken as reflecting some true relationships.

Nonchance expectation. The nonchance expectation (NCE) refers to a number of significant results which is so large that it is unlikely to occur on the basis of chance. For the present purposes, the nonchance expectation is formally defined as the smallest number of significant results such that the chance probability of obtaining that many significant results or more is less than α (the significance level chosen by the investigator). In order to determine the nonchance expectation, it is necessary to know the chance probability distribution for obtaining different numbers of significant results. If the results are independent (the outcome of one result does not influence the outcome of another result), then the chance probability distribution is known and given by the binomial distribution. If the set of results contains dependencies, then the appropriate chance probability distribution is unknown. Thus, the nonchance expectation can only be defined in terms of independent results.

For a set of N independent results, the chance probability--P(S)--of obtaining S significant results is given by the binomial distribution as:

$$P(S) = \frac{N!}{S!(N - S)!} \alpha^{S}(1 - \alpha)^{N-S} .$$

The chance probability of S or more significant results is P(S) + P(S+1) + P(S+2) + ... + P(N). Table 1 presents the chance probabilities for obtaining K or more significant results for a set of 100 independent results and a significance level of .05. According to Table 1, the probability of observing 10 or more significant results is less than α (.05) while the probability of 9 or more significant findings is greater than α. Therefore, 10 is the nonchance expectation for this example. Note that the nonchance expectation depends only

Table 1

Chance Probabilities for Obtaining
K or More Significant Results

K	Probability
0	1.00
1	.99
2	.96
3	.88
4	.74
5	.56
6	.38
7	.23
8	.13
9	.06
10	.02
11	.01
12	.01

Note: This table is based upon 100 independent results with $\alpha = .05$. Probabilities are taken from the binomial distribution.

upon the number (N) of results in a set and the significance level (α). Appendix A presents a table of nonchance expectations for varying N and α.

The nonchance expectation test for a set of results is a sign test involving simple comparison of the number of obtained significant results with the nonchance expectation. If the obtained number is as large as the nonchance expectation, then it is quite likely that some true relationships exist. When no true relationships exist, the obtained number should usually (probability $1 - \alpha$) be less than the nonchance expectation. Thus the following decisions can be made on the basis of the nonchance expectation test.

(1) If the observed number of significant results is less than the nonchance expectation, then the entire set of results is assumed to be chance produced.

(2) If the observed number of significant results is greater than or equal to the nonchance expectation, then the set of results is assumed to reflect some true relationships.

Chance expectation test vs. nonchance expectation test. Which test provides the more appropriate analysis of a set of results? Which test allows the more accurate decision about a set of results? Perhaps a more pertinent question is whether either test is at all appropriate for analyzing a set of correlational results.

Previous suggestion (Godbout, 1975) that the nonchance expectation test be applied to sets of correlational results has met with justified criticism. The nonchance expectation test rests upon the assumption that the results of interest are all independent, a condition never realized in a set of all possible predictor-criterion correlations. Does this preclude practical use of the nonchance expectation test? When applied to independent results, the use of the nonchance expectation test will lead to a Type I error (rejecting the null hypothesis

when it is true) with probability approximately equal to α. What is the Type I error rate when the test is applied to sets of dependent correlational results? What is the statistical power (probability of detecting true, nonchance relationships) when this test is applied to correlational results? Type I error rate and statistical power are the two salient properties determining the usefulness of a statistical test. Without knowledge of the Type I error rate and the statistical power for the nonchance expectation test, the appropriateness of that test is unknown.

The properties of the chance expectation test are also largely unknown. The binomial distribution can be used to determine the Type I error rate and statistical power for the application of the chance expectation test to sets of independent results. However, the Type I error rate and statistical power for sets of dependent correlational results are unknown.

To study the properties of the nonchance expectation test and the chance expectation test in application to sets of correlational results, a series of computer simulations were performed. Sets of dependent correlations were generated. The chance expectation and nonchance expectation tests were then applied to these sets of results. The Type I error rate and statistical power for both tests were determined and compared. The procedures and results of these simulations will now be discussed.

Simulation Procedures

Of interest here was the simulation of sets of all possible predictor-criterion correlations. A set of all possible predictor-criterion correlations always involves dependencies. Let P_1 represent the first predictor variable; P_2, the second predictor variable; and C_1, the first criterion variable. The pair of correlations, $r_{P_1C_1}$ and $r_{P_2C_1}$, are always dependent because they both involve the same criterion variable, i.e., there is a variable common to both correlations. To illustrate dependence due to a common variable, consider the

case in which the correlation between P_1 and P_2 is 0. In this case, if r_{P_1C} equals 1.0, then the value of r_{P_2C} must equal 0. That is, the value of r_{P_2C} is influenced by the value of r_{P_1C}. A nonzero correlation between P_1 and P_2 will provide a second source of dependence between $r_{P_1C_1}$ and $r_{P_2C_1}$. To illustrate dependence due to correlated predictors, consider the case in which P_1 and P_2 are perfectly correlated ($r = 1.0$). In this case, the value of r_{P_2C} is completely determined by r_{P_1C}, these two correlations always equaling the same value. Thus, the dependencies within a set of all possible predictor-criterion correlations arise from two sources: (1) some correlations involve a common predictor variable or a common criterion variable and (2) some correlations may involve different but interrelated predictor variables or criterion variables. The present simulations produced sets of results (predictor-criterion correlations) which involved both sources of dependence.

The plan for each of the present simulations was to generate predictor and criterion scores for samples of subjects by randomly sampling from a multivariate normal population in which (1) all the predictor-criterion correlations were specified and (2) all predictor-predictor and criterion-criterion correlations were specified. The resulting scores were intercorrelated within each sample and the number of significant predictor-criterion correlations was counted for each sample. The number of samples resulting in K significant results (for $K = 0, 1, 2, \ldots, N$, where N is the number of predictor-criterion correlations) was counted. The proportion of samples resulting in K significant results (for $K = 0, 1, 2, \ldots, N$) was calculated and taken as an estimate of the probability of obtaining that many significant results in a sample randomly drawn from the population in question. In this manner, an estimate of the probability distribution for different numbers of observed significant results was obtained for specified populations. Such probability distributions allow estimates of (a) Type I error rates when the population values of all predictor-criterion

correlations equal 0 and (b) statistical power when nonzero population values exist for predictor-criterion correlations.

Score generation was accomplished with Subroutine GGNRM from the International Mathematical and Statistical Libraries (see Reference Note 1). This subroutine generates a sample of multivariate normal deviates (z-scores) for I subjects and J variables. The generated normal deviates are distributed with 0 mean, unit variance, and correlation matrix specified. The algorithm for score generation is the triangular factorization method.[3]

Briefly, for a J variable population correlation matrix Subroutine GGNPM employs the triangular factorization method to arrive at a set of J orthogonal factors such that only Variable 1 loads on the first factor; Variables 1 and 2 load on the second factor; Variables 1, 2 and 3 load on the third factor;...; and all J variables load on the Jth factor. A random number generator is used to produce J factor scores for each of the I subjects in a sample, a task made easy by the orthogonal nature of the factors. Given the triangular factorization results and the randomly generated factor scores, the scores on the original J variables are then determined. Since only the first variable loads on the first factor, a subject's score on the first variable can be calculated from the first factor score for that subject. Since only the first and second variables load on the second factor, a subject's score on Variable 2 can be calculated from the Factor 2 score for that subject and the previously calculated Variable 1 score for that subject. Since Variables 1, 2 and 3 load on Factor 3, a subject's score on the third variable can be calculated from the Factor 3 score for that subject and the previously calculated scores on Variables 1 and 2 for that subject. Generation of scores for the fourth through Jth variables proceeds in similar fashion.

---

[3]Discussion of the triangular factorization method, a well-known mathematical technique, can be found in treatments of factor analysis--for example, see Guertin and Bailey (1970)--and treatments of matrix theory--for example, see Rao (1973). The method is known by several names. Guertin and Bailey's term is diagonal factoring while Rao's term is Gram-Schmidt orthogonalization.

These simulation procedures should be equivalent to randomly sampling scores from a multivariate normal population with specified correlation matrix. To check this supposition sample correlations obtained in the simulations were compared to the population correlations. Appendix B presents these comparisons. Examination of the comparisons in Appendix B indicates that, in general, the simulated correlations varied around the population values and the deviations from the population values appear to indicate an appropriate amount of sampling error. Thus, the present simulation procedures appear to produce valid random samples from a multivariate normal population with specified interrelationships (correlation matrix).

Detailed examination of one of the simulations (Simulation 13C) should serve to clarify the simulation procedures. The population correlation matrix for Simulation 13C is presented in Table 2. Note that the matrix in Table 2 has been divided into several submatrices. The upper left submatrix (the P matrix) involves the correlations among the predictor variables. The lower left submatrix (the P-C matrix) presents all possible predictor-criterion correlations. The lower right submatrix (the C matrix) involves the correlations among the criterion variables. The values in the P matrix and C matrix are actual correlations obtained in an educational research investigation.[4] Note that all predictor-criterion correlations (P-C matrix) are 0, and therefore Simulation 13C is relevant to Type I error rates. Thus, the simulated predictor-criterion correlations were chance-produced deviations from population values of 0, i.e., sets of obtained predictor-criterion correlations were chance-produced sets of results.

---

[4]The entries in a correlation matrix are interrelated in specified ways. In mathematical terms, a correlation matrix is a gramian matrix (Guertin and Bailey, 1970). In order to insure that the matrix for Simulation 13C was a true correlation matrix, correlation values actually obtained by Borich, Godbout, Peck, Kash, and Poynor (1974) were chosen for the entrees in the P matrix and C matrix.

Table 2

Population Intercorrelation Matrix for the 5 Predictor and the 5 Criterion Variables in Simulation 13 C

	Predictor Variables					Criterion Variables				
	P_1	P_2	P_3	P_4	P_5	C_1	C_2	C_3	C_4	C_5
P_1	1.0									
P_2	.1045	1.0								
P_3	.0491	.2144	1.0							
P_4	-.0034	.0487	.1078	1.0						
P_5	-.0080	-.0861	-.0764	-.2487	1.0					
C_1	0	0	0	0	0	1.0				
C_2	0	0	0	0	0	-.4280	1.0			
C_3	0	0	0	0	0	.5653	-.4239	1.0		
C_4	0	0	0	0	0	-.1415	.0596	-.0429	1.0	
C_5	0	0	0	0	0	-.0968	.1391	-.1286	.1871	1.0

Note: The redundant entries above the diagonal have been omitted.

The correlation matrix in Table 2 defines the population relevant to Simulation 13C. This correlation matrix was used to generate 100 samples of scores each involving 72 subjects. A significance level of .10 was chosen with the resulting critical value of r equaling .195. The results of this simulation are presented in Table 3. In Table 3, K represents the number of obtained significant predictor-criterion correlations in a sample; "Observed Frequency," the number of samples in which K significant results were obtained; "Observed Percent," the percentage of samples in which K significant results were obtained; and "Observed Cumulative Percent," the percentage of samples in which K or more significant results were obtained. The final column ("Observed Cumulative Percent") estimates the probability of obtaining K or more significant results in a set of 25 chance-produced results dependent according to the correlation matrix in Table 2.

The results of Simulation 13C were used to estimate Type I error rates for both the chance expectation test and the nonchance expectation test for the population of correlation values in Table 2. The chance expectation is 2.5 since there are predictor-criterion correlations and α is .10. The estimated probability of exceeding the chance expectation, i.e., the probability of obtaining three or more significant results is .50. Fifty percent of the time, the chance expectation test will accept a set of chance-produced results as reflecting true relationships. This is a very high Type I error rate. Inspection of Appendix A indicates that five is the nonchance expectation. The estimated Type I error rate for the nonchance expectation test is .10 since 10% of the samples produced a number of significant results as large as the nonchance expectation. For 100 simulated sets of results with dependence given by the correlation matrix in Table 2, the Type I error rate for the nonchance expectation test is seen to exactly equal α, the Type I risk chosen a priori by the investigator.

Table 3

Results of Simulation 13C
(100 samples, 72 Ss, α = .10)

K	Observed Frequency	Observed Percent	Observed Cumulative Percent
0	4	.04	1.00
1	22	.22	.96
2	24	.24	.74
3	19	.19	.50
4	21	.21	.31
5	3	.03	.10
6	4	.04	.07
7	3	.03	.03

The outcome of Simulation 13C indicates that the nonchance expectation test has an appropriate Type I error rate in application to sets of results defined by the correlation matrix in Table 2. The chance expectation test fares less well. Fifty percent of the time, this test will incorrectly reject the hypothesis that the entire set of results are chance-produced. These conclusions based upon Simulation 13C are not generally true of all the present simulations. Simulation 13C has received special attention here only to provide detailed view of the outcome of an example simulation. This particular simulation was chosen because it was handy at the time of writing. We now turn our attention to the results of the present simulations.

Simulation Results

The reader should be warned that this discussion of the simulation results is quite lengthy and at times tedious. We have tried to provide a wide range of information concerning the feasibility of using the chance expectation and nonchance expectation tests to analyze sets of results. Our aim was to arrive at practical recommendations concerning the use of these two tests. Unfortunately, these practical recommendations could not be formulated on the basis of complete information. While a large number of simulations were performed in the present effort, it was impossible to address all interesting questions concerning the characteristics of the two tests. Thus, our recommendations concerning the use of the two tests are based upon the partial information available at the present time. In order to actually arrive at use recommendations, it has been necessary at times to over-interpret the available simulation results. For example, we have often assumed that relationships found for one or two significance levels would hold for all significance levels. In addition, we have identified trends even when the evidence was somewhat unclear. Such over-interpretation was necessary to arrive at practical recommendations. Less damage is done by such over-interpretation than by ignoring the issue of analyses

of sets of results. Decisions concerning analysis of sets of results must be made whether complete information relevant to such decisions is available at present or not. The analysis recommendations arrived at in the present report are tentative and can be modified as additional information becomes available.

The dependence within a set of predictor-criterion correlations is determined by the population correlation matrix for all predictor and criterion variables. Recall that dependence within a set of predictor-criterion correlations arises from two sources: (1) common variables, i.e., some predictor-criterion correlations have a predictor or a criterion variable in common, and (2) correlated variables, i.e., some predictor-criterion correlations involve different but correlated predictor or criterion variables. An appropriate starting point for our discussion of simulation results is to consider sets of predictor-criterion correlations which involve predictor variables which are uncorrelated with each other and criterion variables which are uncorrelated with each other. Such sets of results are dependent because of common variables but not because of correlated variables. We will refer to such sets of results as minimally dependent.

Minimally dependent sets of results--Type I error rates. Fifteen simulations were performed to study Type I error rates for minimally dependent sets of chance-produced results. The population correlation matrices for each of these simulations had all nondiagonal entrees equal to zero. Table 4 summarizes the outcomes of these simulations. The first four columns of this table describe characteristics of the simulations and the remaining columns present results. "SIM #" refers to codes identifying the simulations. "# Results" is the number of predictor-criterion correlations constituting a set of results. Note that all of the present simulations involved equal numbers of predictor and criterion variables, the number of either type of variable being the square root of the "# Results." In Table 4, the Type I error rate for the CE (chance expectation) test is the proportion of samples obtaining a number of significant results exceeding the chance expectation, i.e., the proportion of samples for which the CE test would reject the true null hypothesis. The Type I error rate for the NCE (nonchance expectation) test is the proportion of samples obtaining a number of significant results either equaling or exceeding the nonchance expectation, i.e., the porportion of samples for which the NCE test would reject the true null hypothesis.

When applied to minimally dependent sets of chance-produced results, the CE test produced quite high Type I error rates. In 11 of the 15 simulations presented in Table 4, the Type I error rate was .4 or higher. That is, in 11 of the 15 cases, a true null hypothesis would be rejected more than 40% of the time. The CE test is clearly quite liberal in rejecting the null hypothesis. In contrast, the NCE test produced much lower Type I error rates than the CE test. In fact, the Type I error rates for the NCE test are quite close to α. The mean Type I error rate across the seven simulations with an α of .05 was .031, while the mean Type I error across the eight simulations with an α of .10

Table 4

Simulations of Minimally Dependent Sets of

Chance-Produced Results

				CE Test		NCE Test	
SIM #	# Results	α	# Samples	CE Value	Type 1 Error Rate	NCE Value	Type 1 Error Rate
17-1A	4	.05	100	.20	.15	2	.00
17-2A	9	.05	100	.45	.45	3	.00
17-3A	16	.05	100	.80	.54	4	.06
17-4A	25	.05	100	1.25	.30	4	.01
17-5A	49	.05	100	2.45	.41	7	.08
17-6A	100	.05	100	5.00	.44	10	.04
22A	196	.05	100	9.80	.51	16	.03
17-1B	4	.10	100	.40	.37	2	.05
17-2B	9	.10	100	.90	.62	3	.05
17-3B	16	.10	100	1.60	.38	4	.04
17-4B	25	.10	100	2.50	.44	6	.10
17-5B	49	.10	100	4.90	.48	8	.07
17-6B	100	.10	100	10.00	.44	15	.06
25	100	.10	500	10.00	.41	15	.07
22B	196	.10	100	19.60	.49	26	.08

Note: All simulations are for samples of 42 subjects.

was .065. The NCE test appears to have a satisfactory Type I error rate in application to minimally dependent sets of results, this test generally yielding a Type I error rate close to but somewhat less than α.

The Type I error rates in Table 4 for both the CE and NCE tests are quite similar to the Type I error rates for these tests as applied to binomially distributed results (i.e., independent results). This suggests the interesting possibility that the chance probability of obtaining different numbers of significant results in a set of minimally dependent results may conform to the binomial distribution. Table 5 presents information relevant to this possibility. In Table 5, the "BIN" columns present Type I error rates for binomially distributed results[5]; the "SIM" columns present Type I error rates obtained in the simulations; and "DIFF" is the binomial Type I error rate minus the simulated Type I error rate. The "DIFF" columns are of main interest. For both the CE test and NCE test, the differences seem to vary unsystematically around zero.

The mean of the 15 values in the "DIFF" column for the CE test is .0118. A *t*-test was used to determine if this mean difference for the CE test was significantly different from zero. This test did not approach significance--the obtained *t* was .929 (*df* = 14) and the *t* required for significance at the .10 level is 1.761. Thus, the binomial and simulated Type I error rates did not significantly differ for the CE test.

The mean of the 15 values in the "DIFF" column for the NCE test is .0031. This mean difference did not significantly differ from zero--the obtained *t* was .660 (*df* = 14) and a value of 1.761 is required for significance at the .10 level. Thus, the binomial and simulated Type I error rates did not significantly differ for the NCE test.

---

[5]An extended-precision routine was used to calculate reported binomial probabilities (to six decimal places) for any α and up to 10,000 events (results). Information concerning this routine is available from Paul T. Marston, Research and Development Center for Teacher Education, The University of Texas at Austin, Austin, Texas 78712.

Table 5

Minimally Dependent Sets of Chance-Produced Results:

Comparison of Type 1 Error Rates Obtained in Simulations

With Error Rates Calculated From the Binomial

SIM #	CE TEST			NCE TEST		
	BIN	SIM	DIFF	BIN	SIM	DIFF
17-1A	.1855	.1500	.0355	.0140	.0000	.0140
17-2A	.3698	.4500	-.0802	.0084	.0000	.0084
17-3A	.5599	.5400	.0199	.0429	.0600	-.0171
17-4A	.3576	.3000	.0576	.0341	.0100	.0241
17-5A	.4463	.4100	.0363	.0347	.0800	-.0453
17-6A	.3840	.4400	-.0560	.0282	.0400	-.0118
22A	.5194	.5100	.0094	.0380	.0300	.0080
17-1B	.3439	.3700	-.0261	.0523	.0500	.0023
17-2B	.6126	.6200	-.0074	.0530	.0500	.0030
17-3B	.4853	.3800	.1053	.0684	.0400	.0284
17-4B	.4629	.4400	.0229	.0980	.1000	-.0020
17-5B	.5503	.4800	.0703	.0519	.0700	.0181
17-6B	.4168	.4400	-.0232	.0726	.0600	.0126
25	.4168	.4100	.0068	.0726	.0720	.0006
22B	.4967	.4900	.0067	.0839	.0800	.0039

The information in Table 5 supports the contention that chance-produced sets of minimally dependent results are binomially distributed. Simulation 25 was examined to provide an additional test of this contention. Goodness of fit tests were performed between the results of Simulation 25 and the binomial distribution. This simulation was chosen for goodness of fit tests because it involved 500 samples and the other simulations of minimally dependent results involved only 100 samples. Table 6 presents frequency and probability comparisons between the results of Simulation 25 and the binomial distribution. A Chi-square test for goodness of fit (McNemar, 1962) was applied to the observed (simulated) and binomial frequencies in Table 6. The resulting Chi-square was 13.969 with 11 degrees of freedom, this value being significant between .20 and .30. A Kolmorgorov-Smirnov (K-S) test for goodness of fit (McNemar, 1962) was also used to compare the fit between the Simulation 25 results and the binomial distribution. The absolute magnitude of the largest cumulative probability difference in Table 6 is .0327 which does not approach the K-S critical value of .0510 required for significance at the .10 level. This lack of significant results for both goodness of fit tests suggests that the results of Simulation 25 are binomially distributed.

In summary, the similarity of binomial and simulated Type I error rates for all 15 simulations of miniamlly dependent results, together with the goodness of fit results for Simulation 25, supports the contention that chance-produced sets of minimally dependent results are binomially distributed. This finding is surprising and useful. It implies that the NCE test--a test based upon the binomial distribution--should be appropriate for application to minimally dependent sets of results. Furthermore, this finding implies that Type I error rates for minimally dependent sets of results need not be simulated, rather these error rates can be determined from the binomial distribution. Table 7 presents

Table 6

Comparison of Simulation 25 With the

Binomial Distribution

# Significant Results	Frequency of Occurrence		Cumulative Probability		
	Observed	Binomial	Observed	Binomial	Difference
0 to 4	11	11.86	1.0000	1.0000	.0000
5	18	16.96	.9780	.9763	.0017
6	33	29.80	.9420	.9424	-.0004
7	41	44.45	.8760	.8828	-.0068
8	66	57.40	.7940	.7939	.0001
9	73	65.20	.6620	.6791	-.0171
10	53	65.95	.5160	.5487	-.0327
11	53	59.90	.4100	.4168	-.0068
12	40	49.40	.3040	.2970	.0070
13	48	37.15	.2240	.1982	.0262
14	28	25.65	.1280	.1239	.0041
15	12	16.35	.0720	.0726	-.0006
16	10	9.65	.0480	.0399	.0081
17+	14	10.30	.0280	.0206	.0074

Type I error rates for sets of binomially distributed results with set sizes varying from 4 to 5,000 results. It can be seen in this table that the NCE test results in a Type I error rate usually quite close to α, the a priori risk of a Type I error chosen by the investigator. The Type I error rates for the CE test are less appropriate, this test producing Type I error rates far exceeding α. In Table 7, the Type I error rate for the CE test is usually in the .3 to .5 range for α equal to either .05 or .10.

Minimally dependent sets of results--Statistical power. The present discussion assumes that the reader has some familiarity with the concept of statistical power. For the unfamiliar reader, Cohen (1969) discusses the power of parametric tests while Bradley (1968) deals with nonparametric tests. The statistical power of a particular test is a complex positive function of the significance level (α), the sample size, and the effect size. With regard to the tests of interest in the present paper (the CE and NCE tests), sample size refers to the number of results in a set of results. A meaningful effect size index for these two tests is the percent of the results reflecting true (non-null) population relationships.

Complex calculations or computer simulations are unnecessary when determining the power of the CE and NCE tests as applied to minimally dependent sets of results. Rather, the fact that minimally dependent sets of chance-produced results appear to be binomially distributed allows approximate power estimates to be calculated from the binomial distribution. For example, consider a set of 10 minimally dependent results. Assume that 2 of these results reflect nonzero correlations in the population while the remaining 8 reflect zero population correlations. Further suppose that the 2 true relationships are strong enough to produce significant sample correlations with probability approaching 1.0. Then, the number of significant results obtained in a sample will be 2

Table 7

Type I Error Rates for Binomially Distributed Results

# Results	α	Type I Error Rate	
		CE Test	NCE Test
4	.05	.1855	.0140
10	.05	.4013	.0115
25	.05	.3576	.0341
50	.05	.4595	.0378
100	.05	.3840	.0282
200	.05	.4169	.0444
500	.05	.4471	.0454
1,000	.05	.4625	.0384
5,000	.05	.4832	.0444
4	.10	.3439	.0523
10	.10	.2639	.0702
25	.10	.4629	.0980
50	.10	.3839	.0579
100	.10	.4168	.0726
200	.10	.4408	.0672
500	.10	.4624	.0810
1,000	.10	.4734	.0954
5,000	.10	.4881	.0981

(the 2 true relationships) plus the number of chance-produced significant results which occur for the other 8 correlations. For the present purposes, the significance of the results for these other 8 correlations (those with corresponding population values of 0) will be assumed to be binomially distributed. The probability of obtaining 2 significant results in the set of 10 is then the binomial probability of obtaining 0 significant results in the set of 8 correlations with corresponding population values of 0. The probability of obtaining 3 significant results in the set of 10 is the binomial probability of obtaining 1 significant result in a set of 8 chance-produced results. The probability of obtaining 4 significant results is the binomial probability of obtaining 2 significant results in a set of 8 chance-produced results. In general, the probability of N significant results in this set of 10 results is the binomial probability of obtaining N-2 significant results in a set of 8 chance-produced results. The probabilities reported in the present section of this paper were determined in similar fashion.

Table 8 presents statistical power values for the CE and NCE tests for α = .05 and for varying numbers of results. Table 9 presents the corresponding information for α = .10. In these tables "# Results" refers to the number of results in a set of results. "# True Relationships" refers to the number of nonzero predictor-criterion relationships in the population. "% of Results Reflecting True Relationships" is 100 times the "# True Relationships" divided by "# Results." "Statistical Power" is the probability of exceeding the chance expectation for the CE test and the probability of equaling or exceeding the nonchance expectation for the NCE test. To insure understanding of the method for determining power values in Tables 8 and 9, let us consider the power values for 25 results in Table 8. For 25 results and α = .05, the chance expectation is 1.25. In order for the CE test to reject the null hypothesis, 2 or more significant results must be obtained in the set of 25. When 1 true

Table 8

Statistical Power of the CE and NCE Tests as a Function of the Number of True Relationships* for $\alpha = .05$ and Minimally Dependent Results

# Results	# True Relationships*	% of Results Reflecting True Relationships*	Statistical Power	
			CE Test	NCE Test
4	1	25.0	1.0000	.1426
	2	50.0	1.0000	1.0000
10	1	10.0	1.0000	.0712
	2	20.0	1.0000	.3366
	3	30.0	1.0000	1.0000
25	1	4.0	.7080	.1159
	2	8.0	1.0000	.3206
	3	12.0	1.0000	.6765
	4	16.0	1.0000	1.0000
50	1	2.0	.7101	.0970
	2	4.0	.9147	.2180
	3	6.0	1.0000	.4195
	4	8.0	1.0000	.6768
	5	10.0	1.0000	.9006
	6	12.0	1.0000	1.0000
100	1	1.0	.5550	.0599
	2	2.0	.7278	.1177
	3	3.0	.8688	.2120
	4	4.0	.9560	.3480
	5	5.0	.9923	.5181
	6	6.0	1.0000	.6972
	7	7.0	1.0000	.8496
	10	10.0	1.0000	1.0000
200	1	0.5	.5389	.0756
	2	1.0	.6614	.1227
	3	1.5	.7728	.1893
	4	2.0	.8633	.2772
	5	2.5	.9280	.3849
	7	3.5	.9882	.6315
	9	4.5	.9994	.8456
	16	8.0	1.0000	1.0000

(Table continued on next page)

*In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

Table 8 (Cont.)

Statistical Power of the CE and NCE Tests as a Function of the Number of True Relationships* for $\alpha = .05$ and Minimally Dependent Results

# Results	# True Relationships*	% of Results Reflecting True Relationships*	Statistical Power	
			CE Test	NCE Test
500	2	0.4	.6022	.0909
	4	0.8	.7464	.1657
	6	1.2	.8600	.2747
	8	1.6	.9348	.4145
	10	2.0	.9751	.5703
	12	2.4	.9924	.7198
	14	2.8	.9982	.8414
	20	4.0	1.0000	.9907
	34	6.8	1.0000	1.0000
1,000	4	0.4	.6782	.1050
	8	0.8	.8501	.2343
	12	1.2	.9484	.4278
	16	1.6	.9876	.6466
	20	2.0	.9980	.8292
	30	3.0	1.0000	.9934
	40	4.0	1.0000	1.0000
	63	6.3	1.0000	1.0000
5,000	10	0.2	.7184	.1352
	20	0.4	.8864	.3103
	30	0.6	.9676	.5480
	40	0.8	.9937	.7718
	50	1.0	.9992	.9160
	70	1.4	1.0000	.9963
	100	2.0	1.0000	1.0000
	277	5.5	1.0000	1.0000

*In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

Table 9

Statistical Power of the CE and NCE Tests as a Function of the Number of True Relationships* for $\alpha = .10$ and Minimally Dependent Results

# Results	# True Relationships*	% of Results Reflecting True Relationships*	Statistical Power	
			CE Test	NCE Test
4	1	25.0	1.0000	.2710
	2	50.0	1.0000	1.0000
10	1	10.0	.6126	.2252
	2	20.0	1.0000	.5695
	3	30.0	1.0000	1.0000
25	1	4.0	.7075	.2143
	2	8.0	.9114	.4080
	3	12.0	1.0000	.6608
	4	16.0	1.0000	.8906
	5	20.0	1.0000	1.0000
50	1	2.0	.5503	.1119
	2	4.0	.7201	.2000
	3	6.0	.8617	.3286
	4	8.0	.9520	.4927
	5	10.0	.9913	.6711
	6	12.0	1.0000	.8296
	9	18.0	1.0000	1.0000
100	1	1.0	.5355	.1167
	2	2.0	.6554	.1791
	3	3.0	.7655	.2618
	4	4.0	.8563	.3642
	5	5.0	.9225	.4816
	6	6.0	.9643	.6051
	8	8.0	.9961	.8250
	15	15.0	1.0000	1.0000
200	1	0.5	.5251	.0961
	2	1.0	.6098	.1339
	3	1.5	.6911	.1816
	4	2.0	.7653	.2397
	5	2.5	.8297	.3079
	10	5.0	.9858	.7203
	12	6.0	.9969	.8532
	27	13.5	1.0000	1.0000

(Table continued on next page)

*In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

Table 9 (Cont.)

Statistical Power of the CE and NCE Tests as a Function of the Number of True Relationships* for $\alpha = .10$ and Minimally Dependent Results

# Results	# True Relationships*	% of Results Reflecting True Relationships*	Statistical Power	
			CE Test	NCE Test
500	1	0.2	.5159	.1017
	3	0.6	.6221	.1546
	5	1.0	.7212	.2242
	7	1.4	.8068	.3100
	9	1.8	.8750	.4092
	11	2.2	.9251	.5160
	13	2.6	.9586	.6233
	15	3.0	.9791	.7232
	20	4.0	.9975	.9047
	60	12.0	1.0000	1.0000
1,000	4	0.4	.6235	.1731
	8	0.8	.7587	.2842
	12	1.2	.8633	.4231
	16	1.6	.9326	.5746
	20	2.0	.9714	.7178
	30	3.0	.9984	.9422
	40	4.0	1.0000	.9956
	113	11.3	1.0000	1.0000
5,000	10	0.2	.6539	.1909
	20	0.4	.7952	.3250
	30	0.6	.8956	.4880
	40	0.8	.9547	.6543
	50	1.0	.9835	.7959
	60	1.2	.9950	.8963
	80	1.6	.9997	.9837
	100	2.0	1.0000	.9988
	529	10.6	1.0000	1.0000

*In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

relationship exists, then the power of the CE test is the probability of obtaining 1 or more significant results for the 24 results not corresponding to the true relationship. The binomial probability for 1 or more significant ($\alpha = .05$) results in a set of 24 results is .7080. In Table 8, this value has been entered as the power of the CE test applied to a set of 25 results with 1 true relationship. When 2 or more true relationships exist, the CE test will always reject the null hypothesis. Thus, power values of 1.0 have been entered in Table 8 for the CE test applied to a set of 25 results with 2, 3, or 4 true relationships.

The nonchance expectation for 25 results and $\alpha = .05$ is given in Appendix A as 4. In order for the NCE test to reject the null hypothesis, 4 or more significant ($\alpha = .05$) results must be obtained in a set of 25. When 1 true relationship exists, then the power of the NCE test is the probability of obtaining 3 or more significant results for the 24 results not corresponding to the true relationship. The binomial probability for 3 or more significant results in a set of 24 results is .1159. In Table 8, this value has been entered as the power of the NCE test as applied to a set of 25 results including 1 true relationship. The second (.3206) and third (.6765) power entrees in Table 8 for the NCE test and 25 results are the binomial probabilities for 2 or more significant results in a set of 23 results and for 1 or more significant results in a set of 22 results. The final NCE power entree for a set of 25 results in Table 8 is for 4 true relationships (i.e., a number of true relationships equal to the nonchance expectation). Since the nonchance expectation will always be equaled or exceeded when there are 4 true relationships, this final power entree is 1.0. Note that the power of the NCE test for 25 results including more than 4 true relationships would also be 1.0.

All power values in Tables 8 and 9 were determined in a fashion similar to that just described for 25 results and $\alpha = .05$. In order to facilitate under-

standing of these two tables, a few words are necessary concerning the tabled effect size measures. The effect sizes in Tables 8 and 9 are labeled as "% of Results Reflecting True Relationships." The true relationships referenced here are strong true relationships. By strong true relationships, we mean population relationships strong enough to produce significant sample results close to 100% of the time--i.e., the probability of a significant sample result, given the population relationship, is close to 1.0.[6] If the effect size entries ("% of Results Reflecting True Relationships") in Tables 8 and 9 are doubled, then the resulting values represent the % of results reflecting marginal true relationships, with a marginal population relationship being defined as one which will produce a significant sample result 50% of the time. The following discussion is couched in terms of strong true relationships. However, the conclusions can be extended to marginal true relationships by doubling the effect size measures ("% of Results Reflecting True Relationships").

As previously mentioned, sample size, significance level and effect size are all positively related to the statistical power of a test (Cohen, 1969; Borich and Godbout, 1974).Positive relationships between power and sample size and between power and effect size are evidenced in Tables 8 and 9. For fixed significance level and fixed effect size, power increases with the number of results in a set of results (sample size). Such a relationship is found throughout Tables 8 and 9 for both the CE test and the NCE test. For example in Table 8, for $\alpha = .05$ and an effect size ("% of Results Reflecting True Relationships") of 4.0, the power of the CE test is .7080 for 25 results but .9147 for 50 results. Similarly, the power of the NCE test, with $\alpha = .05$ and an effect size of 4.0, increases from .1159 for 25 results to .2180 for 50 results. For fixed α and sample size, power increases as effect size increases.

---

[6]Note that the actual numerical value defining a strong population relationship varies with the number of Ss in the sample (i.e., the number of Ss upon which sample correlations are based). The more Ss in the sample, the lower the sample r needed for significance and the lower the population r defining a strong population relationship.

Such a relationship is found throughout Tables 8 and 9 for both the CE test and the NCE test. For example in Table 8, for $\alpha = .05$ and a sample size ("# Results") of 50, the power of the CE test is .7101 for an effect size of 2.0 but .9147 for an effect size of 4.0. Similarly, the power of the NCE test, with $\alpha = .05$ and a sample size of 50, increases from .0970 for an effect size of 2.0 to .2180 for an effect size of 4.0.

While power and significance level are positively related for statistical tests in general (Cohen, 1969), such a relationship is not evidenced in Tables 8 and 9 for the CE and NCE tests. In Tables 8 and 9, for fixed sample size and effect size, CE test power is generally higher for $\alpha = .05$ than for $\alpha = .10$ -- i.e., power is negatively related to significance level. In these tables, for fixed sample size and effect size, NCE test power can be seen to at times be higher for $\alpha = .10$ than for $\alpha = .05$ but at other times be higher for $\alpha = .05$ than for $\alpha = .10$. Fortunately, there is an explanation for these departures from the normal positive relationship between power and significance level. Both the CE and NCE tests are unlike other statistical tests in that the data to which they are applied vary with significance level. When $\alpha = .05$, the data analyzed with both the CE and NCE tests is the number of results significant at the .05 level. In contrast, when $\alpha = .10$, the data analyzed with both tests is the number of results significant at the .10 level. The fact that different data are involved for different significance levels precludes examination of the CE and NCE tests with regard to the typical relationship between power and significance level.

Tables 8 and 9 also illustrate the relative power of the CE test as compared to the NCE test. The fact that the CE test tends to have greater power than the NCE test can be seen throughout Tables 8 and 9. In fact, for any application involving less than perfect (1.0) power for the NCE test, the statistical power

of the CE test always exceeds that of the NCE test. For small effect sizes, the superiority of the CE test is greatest and, as effect size increases, the superiority of the CE test diminishes. For large enough effect sizes, both tests achieve power of 1.0--i.e., the superiority of the CE test with regard to statistical power completely disappears. Differences in the power of the CE and NCE test also vary with sample size (number of results in a set of results). For a fixed effect size, the power superiority of the CE test decreases with increases in sample size. For example in Table 8, for 50 results and an effect size of 2.0, the power of the CE test is .7101 and the power of the NCE test is .0970 (a difference of .6131). The power values in Table 8 for 5000 results and an effect size of 2.0 are 1.0000 and 1.0000 for the CE test and NCE test, respectively. Thus, a power difference of .6131 for 50 results declines to 0 for 5000 results. This concludes discussion of Table 8 and 9. Although discussion of these tables has been relatively brief, these tables do provide the interested reader with a great deal of information about the power of the CE and NCE tests in application to minimally dependent sets of results.

For which applications does the CE test have high power and for which applications does the NCE test have high power? This is the basic issue of interest in the present section of this paper. To ease discussion of this issue, we will arbitrarily define high power as a probability greater than 0.80 of detecting a proportion of strong true relationships equal to α or less.[7] That

[7] At first glance, it may appear unusual to allow the effect size associated with high power to vary with α. However, recall that the data analyzed with the CE and NCE tests vary with α. When $\alpha = .05$, these tests involve analysis of the number of results significant at the .05 level while, for $\alpha = .10$, they involve analysis of the number of results significant at the .10 level. Defining high power in terms of a constant effect size for differing significance levels would therefore seem inappropriate. Certainly, 5% of the results significant at the .05 level represents a stronger effect than 5% of the results significant at the .10 level. That is, an effect size of 5.0 with $\alpha = .05$ is not comparable to an effect size of 5.0 with $\alpha = .10$. It would seem more reasonable to assume that an effect size of 5.0 with $\alpha = .05$ is comparable in terms of strength of effect to an effect size of 10.0 with $\alpha = .10$. The present definition of high power rests on such an assumption. Further rationale for this assumption will be offered later in the present section.

is, for $\alpha = .05$, high power is achieved if the probability of detecting 5% or fewer strong true relationships is greater than 0.80. For $\alpha = .10$, high power is a probability exceeding 0.80 of detecting 10% or fewer strong true relationships. We will also arbitrarily define very high power as a probability greater than 0.80 of detecting a proportion of strong true relationships equal to one-half α or less. That is, for $\alpha = .05$, very high power consists of a probability exceeding 0.80 of detecting 2.5% or fewer strong true relationships. For $\alpha = .10$, very high power is achieved if 5% or fewer strong true relationships are detected with probability exceeding 0.80.

The information in Tables 8 and 9 allows determination of the applications for which the CE and NCE tests have high power and very high power. However, such determination is greatly facilitated if information from Tables 8 and 9 is reorganized as in Table 10. Table 10 gives effect size values ("% of Results Reflecting Strong True Relationships") required for power to exceed 0.80. The first row of Table 10 is for a sample size ("#Results") of 4 with subsequent rows being for larger sample sizes. The first 3 effect size values in a row are for $\alpha = .05$ and the second 3 effect size values in a row are corresponding values for $\alpha = .10$. The initial two effect size entrees in the first row of Table 10 are 25.0 and 50.0. These values indicate that, for $\alpha = .05$ and a set of 4 results, at least 25% of the results must reflect strong true relationships for the CE test to have power exceeding 0.80, while at least 50% of the results must reflect strong true relationships for the NCE test to have power exceeding 0.80. The third effect size entree in the first row of Table 10 is 0.50. This value represents the ratio of the preceeding CE effect size value (25.0) to the preceeding NCE effect size value (50.0).

The values in Table 10 were arrived at on the basis of calculations analogous to those used for Tables 8 and 9. That is, the calculations underlying Table 10 are based on the assumption that chance-produced sets of

Table 10

Minimum Effect Sizes[a] Required for Power > 0.80

for Minimally Dependent Results

# Results	% Results Reflecting True Relationships[b] for Power > 0.80					
	$\alpha = .05$			$\alpha = .10$		
	CE Test	NCE Test	CE to NCE Ratio	CE Test	NCE Test	CE to NCE Ratio
4	25.0	50.0	.50	25.0	50.0	.50
10	10.0	30.0	.33	20.0	30.0	.67
25	8.0	16.0	.50	8.0	16.0	.50
50	4.0	10.0	.40	6.0	12.0	.50
100	3.0	7.0	.43	4.0	8.0	.50
200	2.0	4.5	.44	2.5	6.0	.42
500	1.0	2.8	.36	1.4	3.4	.41
1000	0.7	2.0	.35	1.0	2.3	.43
5000	0.3	0.8	.38	0.4	1.0	.40

[a]Effect size refers to the percent of results reflecting true relationships.

[b]In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

minimally dependent results are binomially distributed. Furthermore, these calculations are based upon the assumption that true population relationships are strong enough to always produce a significant sample result. Thus, the effect size values in Table 10 are appropriate for strong true population relationships. Doubling these tabled effect size values will yield the effect sizes appropriate for marginal true population relationships. Recall that a marginal population relationship is one which will produce a significant sample result 50% of the time.

Consider Table 10 with regard to determining when the CE and NCE tests exhibit high power and very high power. When $\alpha = .05$, the CE test has high power (probability greater than 0.80 of detecting 5% strong true relationships) when applied to sets of 50 or more results and very high power (probability greater than 0.80 of detecting 2.5% strong true relationships) when applied to sets of 200 or more results. In contrast, when $\alpha = .05$, the NCE test has high power for sets of 200 or more results and very high power for sets of 1000 or more results. When $\alpha = .10$, the CE test has high power (probability greater than 0.80 of detecting 10% strong true relationships) for sets of 25 or more results and very high power (probability greater than 0.80 of detecting 5% strong true relationships) for sets of 100 or more results. For $\alpha = .10$, 100 or more results are required for the NCE test to achieve high power while 500 or more results are required for this test to attain very high power.

For the present purposes, we will recommend that a test be applied only when it achieves high power or better. While this recommendation is based upon an arbitrary standard, it does not appear to be unreasonable for exploratory educational research. Table 11 presents use recommendations for the CE and NCE tests as applied to sets of minimally dependent results. These recommendations in Table 11 are based upon a test's achieving high power or better. As can be seen in this table, for $\alpha = .05$, neither test is recommended for less than 50

Table 11

Recommended Use of the CE and NCE Tests
on Basis of Achieving High Power*

# Results	Test(s) Recommended for $\alpha = .05$	Tests Recommended for $\alpha = .10$
less than 25	neither CE nor NCE	neither CE nor NCE
25	neither CE nor NCE	CE only
50	CE only	CE only
100	CE only	both CE and NCE
200 or more	both CE and NCE	both CE and NCE

*High power is a probability greater than 0.80 of detecting 5% true relationships for $\alpha = .05$ and 10% true relationships for $\alpha = .10$.

results, the CE test is recommended for 50 or more results, and the NCE test is recommended for 200 or more results. For $\alpha = .10$, neither test is recommended for less than 25 results, the CE test is recommended for 25 or more results, and the NCE test is recommended for 100 or more results.

The recommendations in Table 11 are based upon the present definition of high power and upon acceptance of such a level of power as a minimum standard. It is possible that some readers may be interested in higher power standards than the one adopted for the present purposes. Therefore, tables presenting effect sizes required for power exceeding 0.90 (Table 12), for power exceeding 0.95 (Table 13), and for power equaling 1.00 (Table 14) have been included in the present paper. These three tables follow a format identical to Table 10. If any aspect of the formats of Tables 12 - 14 is unclear, the previous discussion of Table 10 should be consulted.

The general relationships in Tables 12 - 14 are consistent with those already discussed with regard to Tables 8 - 9. However, Table 14 does provide one new piece of information that should be noted. Table 14 demonstrates that the minimum effect size required for perfect (1.00) power becomes asymptotic as sample size becomes very large. For $\alpha = .05$, the effect size required for perfect power approaches 5.00 (i.e., 5% strong true relationships) as an asymptote for both the CE and NCE tests. That is, the minimum effect size which can be detected with perfect power is 5% when $\alpha = .05$. For $\alpha = .10$, the effect size required for perfect power approaches 10.00--i.e., 10% strong true relationships--as an asymptote for both the CE and NCE tests. That is, the minimum effect size which can be detected with perfect power is 10% when $\alpha = .10$. In general, it is our guess that, in order to achieve perfect power, a minimum proportion of strong true relationships equal to α is required. Furthermore, since 5.0 is the minimum effect size detectable with perfect power for $\alpha = .05$ and 10.0 is the minimum effect size detectable

Table 12

Minimum Effect Sizes[a] Required for Power > 0.90
for Minimally Dependent Results

# Results	% Results Reflecting True Relationships[b] for Power > 0.90					
	$\alpha = .05$			$\alpha = .10$		
	CE Test	NCE Test	CE to NCE Ratio	CE Test	NCE Test	CE to NCE Ratio
4	25.0	50.0	.50	25.0	50.0	.50
10	10.0	30.0	.33	20.0	30.0	.67
25	8.0	16.0	.50	8.0	20.0	.40
50	4.0	10.0	.40	8.0	14.0	.57
100	4.0	8.0	.50	5.0	9.0	.56
200	2.5	5.0	.50	3.5	6.5	.54
500	1.4	3.2	.44	2.0	4.0	.50
1000	1.0	2.3	.43	1.4	2.8	.50
5000	0.4	1.0	.40	0.6	1.2	.50

[a]Effect size refers to the percent of results reflecting true relationships.

[b]In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

Table 13

Minimum Effect Sizes[a] Required for Power > 0.95
for Minimally Dependent Results

# Results	% Results Reflecting True Relationships[b] for Power > 0.95					
	$\alpha = .05$			$\alpha = .10$		
	CE Test	NCE Test	CE to NCE Ratio	CE Test	NCE Test	CE to NCE Ratio
4	25.0	50.0	.50	25.0	50.0	.50
10	10.0	30.0	.33	20.0	30.0	.67
25	8.0	16.0	.50	12.0	20.0	.60
50	6.0	12.0	.50	8.0	16.0	.50
100	4.0	9.0	.44	6.0	10.0	.60
200	3.0	5.5	.54	4.0	7.5	.53
500	1.8	3.4	.53	2.6	4.6	.56
1000	1.3	2.5	.52	1.8	3.1	.58
5000	0.5	1.1	.45	0.8	1.4	.57

[a] Effect size refers to the percent of results reflecting true relationships.

[b] In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

Table 14

Minimum Effect Sizes[a] Required for Power = 1.00
for Minimally Dependent Results

# Results	% Results Reflecting True Relationships[b] for Power = 1.00					
	α = .05			α = .10		
	CE Test	NCE Test	CE to NCE Ratio	CE Test	NCE Test	CE to NCE Ratio
4	25.0	50.0	.50	25.0	50.0	.50
10	10.0	30.0	.33	20.0	30.0	.67
25	8.0	16.0	.50	12.0	20.0	.60
50	6.0	12.0	.50	12.0	18.0	.67
100	6.0	10.0	.60	11.0	15.0	.73
200	5.5	8.0	.69	10.5	13.5	.78
500	5.2	6.8	.76	10.2	12.0	.85
1000	5.1	6.3	.81	10.1	11.3	.89
3000	5.0	5.7	.88	10.0	10.7	.93
5000	5.0	5.5	.91	10.0	10.6	.94
10000	5.0	5.4	.92	10.0	10.4	.96

[a]Effect size refers to the percent of results reflecting true relationships.

[b]In this table, a true relationship is assumed to be strong enough to always produce a significant sample result.

with perfect power when $\alpha = .10$, it would seem reasonable to assume that 5.0 with $\alpha = .05$ is an effect size comparable to 10.0 with $\alpha = .10$. The assumed comparability of an effect size of 5.0 with $\alpha = .05$ and an effect size of 10.0 with $\alpha = .10$ has been previously used in forming the present definition for high power. Recall that high power was defined as a probability greater than 0.80 of detecting 5% or fewer strong true relationships when $\alpha = .05$ and of detecting 10% or fewer strong true relationships when $\alpha = .10$.

This concludes discussion of the statistical power of the CE and NCE tests for sets of minimally dependent results. The following section represents an attempt to determine the applications (involving minimally dependent results) for which the CE test is better than the NCE test and vice versa. Determination of which test is better for a given set of results demands simultaneous consideration of Type I error rate and statistical power. Thus, the following section is an attempt to integrate the major findings of the present section on statistical power with the major findings of the preceding section on Type I error rate.

Minimally dependent sets of results--summary and conclusions. In any hypothesis testing endeavor, the ideal is to apply a statistical test or method which simultaneously yields high statistical power and a low Type I error rate. High statistical power allows a test to detect true relationships or effects when they exist, while low Type I error rate allows the detection of a lack of true relationships or effects when they do not exist. An ideal statistical procedure would always yield significant results when true effects are present in the population--i.e., power = 1.0--but never produce significant results when there are no true effects in the population--i.e., Type I error rate = 0. In practical research involving analysis of a sample drawn from a population of interest, such an ideal procedure never exists--power equal to

1.0 and Type I error rate equal to 0 cannot simultaneously be achieved. Moreover, in practical applications, Type I error rate and statistical power tend to be positively related (McNemar, 1962, p. 67). Methods yielding the highest power often yield the highest Type I error rate. Therefore, decisions about what procedure is most appropriate often depend upon which favorable characteristic--high power or low Type I error rate--is more important in the particular application in question. In the present case, where the research of interest is of an exploratory nature, high power would seem to be more important than low Type I error rate. In exploratory research, the intent is to identify possible effects which will then be studied in more detail in follow-up research. Given this situation, it is certainly more advisable to err in the direction of identifying too many possible effects for further study than to err in the direction of missing true relationships and thereby precluding their further study.[8]

While high power is a primary concern in analyzing the results of exploratory research, it is not the sole concern. Consider the case of two statistical tests with the following characteristics.

	Power	Type I error rate
Test A	.90	.90
Test B	.90	.05

Use of Test A is questionable. While Test A has high power, it also has a very high Type I error rate. Finding a significant result with Test A actually provides very little information since a significant result is likely to be obtained regardless of whether true population relationships exist or

---

[8]Brophy and Evertson (1974b, p. 15) make an analogous argument for exploratory research.

not. Test B seems to offer a much more satisfactory approach because of its low Type I error rate. The advantage of Test B over Test A can be summarized as follows. Both tests are likely to lead to a correct conclusion when a true population relationship actually exists. However, when no true relationship exists, Test B will usually provide a correct conclusion while Test A will usually provide an incorrect conclusion. Thus, if two tests both achieve high power, then the test with the lower Type I error rate is to be preferred.

In exploratory research, the primary concern in selecting an analysis method should be a high level of statistical power while a secondary concern is a low Type I error rate. These two concerns and their priorities provided the basic scheme for constructing Table 15. Table 15 presents the methods recommended for analysis of different sets of minimally dependent results. For $\alpha = .05$, no test is recommended for less than 50 results, the CE test is recommended for 50 or more but less than 200 results, and the NCE test is recommended for 200 or more results. For $\alpha = .10$, no test is recommended for less than 25 results, the CE test is recommended for 25 or more but less than 100 results, and the NCE test is recommended for 100 or more results.

The recommendations in Table 15 were arrived at in the following manner. A test has been recommended only when it achieves a high level of power. The standard for high power is that defined in the preceding section of this paper --i.e., high power is a probability greater than 0.80 of detecting a proportion of strong true relationships equal to α or less. The different sets of minimally dependent results for which the CE and NCE tests attain this high power standard have been previously presented in Table 11. If neither the CE test nor the NCE test achieve high power for a particular set of results, then no test has been recommended for that set of results in Table 15. For example, for 25 results and $\alpha = .05$, neither test achieves high power (see Table 11) and therefore neither has been recommended for this case in Table 15.

Table 15

Analysis Methods Recommended as Appropriate
for Different Sets of Minimally Dependent Results

# Results	Test Recommended for $\alpha = .05$	Test Recommended for $\alpha = .10$
less than 25	no test	no test
25	no test	CE test
50	CE test	CE test
100	CE test	NCE test
200 or more	NCE test	NCE test

For some sets of results (e.g., 50 results and $\alpha = .05$), the CE test achieves high power but the NCE test does not--see Table 11. For such sets of results, the CE test is the clear choice and this test has been recommended for such cases in Table 15. Finally, for some sets of results (e.g., 200 results with $\alpha = .05$), Table 11 reveals that both the CE and NCE tests achieve the present standard for high power. In such cases, the NCE test has been recommended since this test has a much lower Type I error rate than the CE test. Recall from the previous discussion of Type I error rates for minimally dependent results that the NCE test usually has a Type I error rate close to α while the CE test usually has a Type I error rate in the range of .40 or more.

The priority given to high power in recommending analysis methods is quite evident in Table 15. For less than 50 results with $\alpha = .05$ and less than 25 results with $\alpha = .10$, no test is recommended. Applying no test is essentially equivalent to accepting any significant results as valid--a conclusion which will frequently be in error. Recall from the introduction to this paper that there is a clear need for overall analysis of sets of results. Recommending no test for certain sets of results runs counter to this need. However, application of either the CE test or the NCE test to these sets of results would produce too great a risk of missing true population relationships--i.e., power is too low. It is our contention that for exploratory research it is better to apply no test and therefore preclude identification of chance-produced sets of results than to undertake such identification when it will produce too high a risk of missing valid population relationships. Not applying any test, while not an optimal alternative, is better than applying a test which does not have high power.

Table 15 recommends the CE test for a moderate number of results (50 or more but less than 200 results for $\alpha = .05$ and 25 or more but less than 100

results for $\alpha = .10$). The CE test is not an optimal statistical test in that it possesses a quite high Type I error rate. Most widely accepted statistical tests have a Type I error rate approximately equal to α. When α is .05 or .10, the CE test has a Type I error rate usually equal to .40 or more. Even with this high Type I error rate, the CE test is an improvement over not applying any test. Furthermore, it is preferable to apply the CE test with its high Type I error rate but low power. The NCE test with its appropriately low Type I error rate is only recommended for relatively large sets of results (200 or more results for $\alpha = .05$ and 100 or more results for $\alpha = .10$). Only for such large sets of results does that test achieve our standard for high power.

Power standards other than the present one could be used in forming use recommendations similar to those in Table 15. Researchers who are dissatisfied with the present standard are free to arrive at their own power standard and their own use recommendations. The present power tables (especially Table 10 and Tables 12-14) will aid in determining use recommendations based on different power standards. The effect of a particular power standard is to establish the lower bound for recommendation of a test--i.e., the minimum number of results for which a test is recommended. For example, with the present minimum power standard (probability exceeding 0.80 of detecting a proportion of strong true relationships equal to α or less), the CE test achieves that standard when applied to 50 or more results with $\alpha = .05$. When $\alpha = .05$, 50 results is the lower bound[9]

[9]Note that 50 is not an exact lower bound but rather an approximate lower bound. Power calculations were not performed for all possible numbers of results. For example, calculations were performed for 25 results and 50 results but not for set sizes between 25 and 50. The recommended lower bounds (Table 15) were selected from set sizes for which calculations were actually performed. The true lower bound will usually be somewhat lower than that recommended in Table 15. This approximate nature of the recommended lower bounds does not pose a real problem. The effect of using approximate lower bounds in forming use recommendations for the CE and NCE tests is simply to make the actual underlying power standard somewhat higher than that described.

for recommended use of the CE test. Use of a power standard different from the present one would serve to change the lower bound for recommending a test. If a lower power standard is adopted, then the lower bound for recommendation of a test would be lowered. If the standard of power is raised, then the lower bound for recommended use would be raised.

It is our contention that the present power standard is reasonable for exploratory research in education. Note that the present power standard represents the minimum power achieved by a recommended test. Except at the lower bound, the power actually achieved by a recommended test will exceed the minimum standard. For example, the lower bound for recommended use of the NCE test is 200 when $\alpha = .05$. At this lower bound, the power of the NCE test is slightly higher than the minimum standard. For 200 results and $\alpha = .05$, the NCE test will detect 4.5% strong true relationships with a probability of .8456 (from Table 8). Since power increases with sample size (number of results), the power of the NCE test increases above this level as the number of results increases above 200. For 500 results and $\alpha = .05$, NCE test power is .9907 for detecting 4.0% strong true relationships (from Table 8). For 5000 results and $\alpha = .05$, the power of the NCE test rounds to 1.0000 for detecting only 2% strong true relationships (from Table 8). Thus, for most recommended applications, NCE test power far exceeds the minimum standard. The same is also true of the CE test.

Even at its minimum for recommended applications, the power of the CE and NCE tests is quite high. The present minimum power standard has been defined as a probability exceeding 0.80 of detecting 5% or fewer strong true

relationships for $\alpha = .05$ and 10% or fewer strong true relatioships for $\alpha = .10$. When considered in isolation, this definition may give too low an impression of the minimum power achieved in recommended applications. The minimum power for recommended applications can be viewed from other perspectives. For example, in recommended applications involving minimum power, what effect sizes are detectable with almost complete certainty (probability exceeding 0.95)? The following power values are taken from Table 13:

a) at the CE test lower bound (50 results) for $\alpha = .05$, 6% strong true relationships will be detected with probability exceeding 0.95;

b) at the NCE test lower bound (200 results) for $\alpha = .05$, 5.5% strong true relationships will be detected with probability exceeding 0.95;

c) at the CE test lower bound (25 results) for $\alpha = .10$, 12% strong true relationships will be detected with probability exceeding 0.95;

d) at the NCE test lower bound (100 results) for $\alpha = .10$, 10% strong true relationships will be detected with probability exceeding 0.95.

From this perspective, effect sizes slightly higher than 5% for $\alpha = .05$ and 10% for $\alpha = .10$ will be detected with almost complete certainty in any recommended application.

In recommended applications involving minimum power what effect sizes are detectable with moderately high probability (probability in the range of 0.6 to 0.7)? The following power values are taken from Tables 8 and 9:

a) at the CE test lower bound (50 results) for $\alpha = .05$, 2% strong true relationships will be detected with probability equal to .7101;

b) at the NCE test lower bound (200 results) for $\alpha = .05$, 3.5% strong true relationships will be detected with probability equal to .6315;

c) at the CE test lower bound (25 results) for $\alpha = .10$, 4% strong true relationships will be detected with probability equal to .7075;

d) at the NCE test lower bound (100 results) for $\alpha = .10$, 6% strong true relationships will be detected with probability equal to .6051.

Thus, in any recommended application, there is at least a moderately high probability (around 0.6 to 0.7) of detecting a proportion of strong true relationships approximately equal to one-half α. This perspective and the preceding one support the contention that the minimum power achieved in recommended applications of the CE and NCE tests is reasonably high.

The analysis recommendations in Table 15 are quite important. Not only do we suggest that these recommendations are appropriate for sets of minimally dependent results, but also we will later present arguments that they are useful with regard to any set of results regardless of the type of dependence among the results. This concludes the discussion of minimally dependent results. Sets of minimally dependent predictor-criterion correlations are rarely found in actual research. More usually, predictor-criterion correlations will involve both predictor variables which are intercorrelated and criterion variables which are intercorrelated. We now turn our attention to analysis of sets of results which are not minimally dependent.

Sets of results involving varying dependence--Type I error rate. In the present section, interest is focused on sets of predictor-criterion correlations involving predictor variables which are correlated with other predictor variables and criterion variables which are correlated with other criterion variables. For the present purposes, such sets of predictor-criterion correlations will be

termed sets of results displaying correlated dependence. A series of simulations were performed to study the Type I error rates for sets of results involving varying degrees of correlated dependence. These simulations followed the procedures previously described in the "Simulation Procedures" section of the present paper.

An example population intercorrelation matrix for a Type I error rate simulation (Simulation 1) involving a set of results with correlated dependence is presented in Table 16. Note that the population intercorrelation matrix has been divided into submatrices. The upper left submatrix (the P matrix) presents intercorrelations among the predictor variables, the lower left submatrix (the P-C matrix) presents the predictor-criterion correlations, and the lower right submatrix (the C matrix) presents the intercorrelations among the criterion variables. Note that there is a uniform intercorrelation of .2 among all pairs of predictor variables (P matrix) and among all pairs of criterion variables (C matrix). Sets of results based upon this population intercorrelation matrix therefore display uniform correlated dependence. In Table 16, note that all population values for predictor-criterion correlations (P-C matrix) are 0. Thus, sets of simulated predictor-criterion results based upon this population matrix will be relevant to Type I error rate--i.e., simulated sets of predictor-criterion correlations will be sets of chance-produced results.

Table 17 presents the results for simulations of Type I error rates for sets of results involving uniform correlated dependence. The population intercorrelation matrices for these simulations follow the pattern presented for Simulation 1 in Table 16. In Table 17, "SIM #" refers to a code identifying the simulations; "# Results" refers to the number of predictor-criterion correlations in a set of results; and "Dependence" is the uniform correlation

Table 16

Population Intercorrelation Matrix for the
3 Predictor and 3 Criterion Variables in Simulation 1

	Predictors			Criteria		
	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.2	1.0				
P_3	.2	.2	1.0			
C_1	0	0	0	1.0		
C_2	0	0	0	.2	1.0	
C_3	0	0	0	.2	.2	1.0

Note: The redundant entries above the diagonal have been omitted.

Table 17

Simulations of Sets of Chance-Produced Results
Involving Uniform Correlated Dependence

SIM #	# Results	Dependence*	Type I Error Rate	
			CE Test	NCE Test
17-2A	9	0	.45	.00
1	9	.2	.39	.02
2	9	.4	.36	.02
3	9	.6	.32	.03
4	9	.8	.30	.04
17-4A	25	0	.30	.01
5	25	.2	.26	.02
6	25	.4	.30	.06
7	25	.6	.22	.06
8	25	.8	.27	.13
17-5A	49	0	.41	.08
9	49	.2	.44	.06
10	49	.4	.34	.13
11	49	.6	.33	.16
12	49	.8	.29	.11
17-6A	100	0	.44	.04
18B	100	.2	.39	.07
20A	100	.4	.32	.18
20B	100	.6	.21	.10
20C	100	.8	.31	.17
22A	196	0	.51	.03
23A	196	.2	.47	.11
23B	196	.4	.43	.26
23C	196	.6	.35	.23
23D	196	.8	.25	.16

Note: All simulations involve $\alpha = .05$, 100 samples and 42 Ss per sample.

*The "Dependence" value is the population correlation for all possible pairs of predictor variables and for all possible pairs of criterion variables.

between all pairs of predictor variables and all pairs of criterion variables. In this table, the Type I error rate for the CE test is the proportion of samples obtaining a number of significant results exceeding the chance expectation--i.e., the proportion of samples for which the CE test would reject the true null hypothesis. The Type I error rate for the NCE test is the proportion of samples obtaining a number of significant results either equaling or exceeding the nonchance expectation--i.e., the proportion of samples for which the NCE test would reject the true null hypothesis. Note that all simulations in Table 17 involved $\alpha = .05$, 100 samples, and 42 subjects per sample. All simulations in this table also involved an equal number of predictor variables and interim variables, the number of either being the square root of the number of results.

Consider the CE test Type I error rates in Table 17. For simulations involving 9 results, 49 results, 100 results and 196 results, an increase in dependence produces a clear decrease in the CE test Type I error rate. For these simulations, 0 dependence produced CE test Type I error rates ranging from .41 to .51, while .8 dependence produced CE test Type I error rates ranging from .25 to .31. Note that such a clear decrease in Type I error rates does not occur for the simulations involving 25 results in Table 17. Although the reasons for this discrepancy are not well understood, it may merely be a reflection of sampling error.

When correlated dependence is 1.0, then the CE test Type I error rate will be α for sets involving any number of results. This fact can be demonstrated as follows. When correlated dependence is 1.0, then it is as if there were only a single predictor variable and a single criterion variable (all predictor variables are perfectly intercorrelated and all criterion variables are perfectly intercorrelated). Thus when dependence is 1.0, all of the predictor-criterion correlations in a set of results will be identical. In this case, the results in a set will either all be significant

or all be nonsignificant. The probability that a single result is significant by chance is α. Therefore, all of the results in a set with 1.0 dependence will be significant by chance with probability equal to α. For sets of results with dependence equal to 1.0, the CE test will reject the null hypothesis only when all results are significant, since the chance expectation will be exceeded only when all results are significant. Since the probability of obtaining all results significant by chance is α, the CE test Type I error rate for 1.0 dependence is also α.

Reconsider the CE test Type I error rates in Table 17. The CE test Type I error rates are in the range of .25 to .31 for dependence equal to .8. As dependence increases from .8 to 1.0, the CE test Type I error rates must decrease from these values to a value of .05 (α = .05). In general then, as dependence increases from 0 to 1.0, the CE test Type I error rate decreases from .4 or higher to .05. Note in Table 17 that the speed of this decrease is relatively constant for different set sizes ("# Results").

A quite different pattern is evidenced in Table 17 for the NCE test Type I error rate. As dependence increases above 0, there is a clear increase in the NCE test Type I error rate. For 0 dependence, the NCE test Type I error rate is usually below α (α = .05 for Table 17) but, for .8 dependence, it has usually increased above α. When dependence equals 1.0, the NCE test Type I error rate can be shown to be equal to α. (The demonstration that the NCE test Type I error rate equals α when dependence equals 1.0 is identical to the previous demonstration that the CE test Type I error rate equals α when dependence is 1.0). In general then, the relationship between NCE test Type I error rate and dependence is nonmonotonic. As dependence increases from 0 to 1.0, the NCE test Type I error rate initially increases from a value slightly below α to a maximum value usually exceeding α and then decreases from this maximum back down to α. Note in Table 17 that the speed and extent of the initial increase in NCE test Type I

error rate seem to vary in orderly fashion with set size ("# Results"). Both the speed and extent of this increase appear to be greater for the larger set sizes than for the smaller set sizes.

Figure 1 is based upon the Type I error rate information in Table 17. This figure presents idealized relationships between dependence and NCE test Type I error rate, as well as idealized relationships between dependence and CE test Type I error rate. Note in this figure that, at 0 dependence (i.e., for minimally dependent results), the NCE test has a much lower Type I error rate than the CE test. In contrast, when dependence equals 1.0, there is no difference between the Type I error rates for the two tests, the error rates for both tests euqaling α. Thus, as dependence increases from 0 to 1.0, a large initial difference between the two tests completely disappears. Furthermore, this initial difference decreases more quickly when there are a large number of results than when there are a small number of results.

Thes relationships in Figure 1 are quite important. They represent our conclusions concerning the effect of uniform correlated dependence on Type I error rate. While the relationships in Figure 1 have been derived from simulations which were limited to $\alpha = .05$ and set sizes of 196 and smaller (i.e., the simulations in Table 17), we find no reason to assume that similar relationships do not generally hold for any α or any set size.

An additional point about Figure 1 should be discussed. The Type I error rate of the NCE test is shown to never exceed that of the CE test. The Type I error rate for the NCE test can never exceed that for the CE test because the number of significant results required for rejection of the null hypothesis is always greater for the NCE test than for the CE test. Recall that the NCE test rejects the null hypothesis whenever the number of significant results

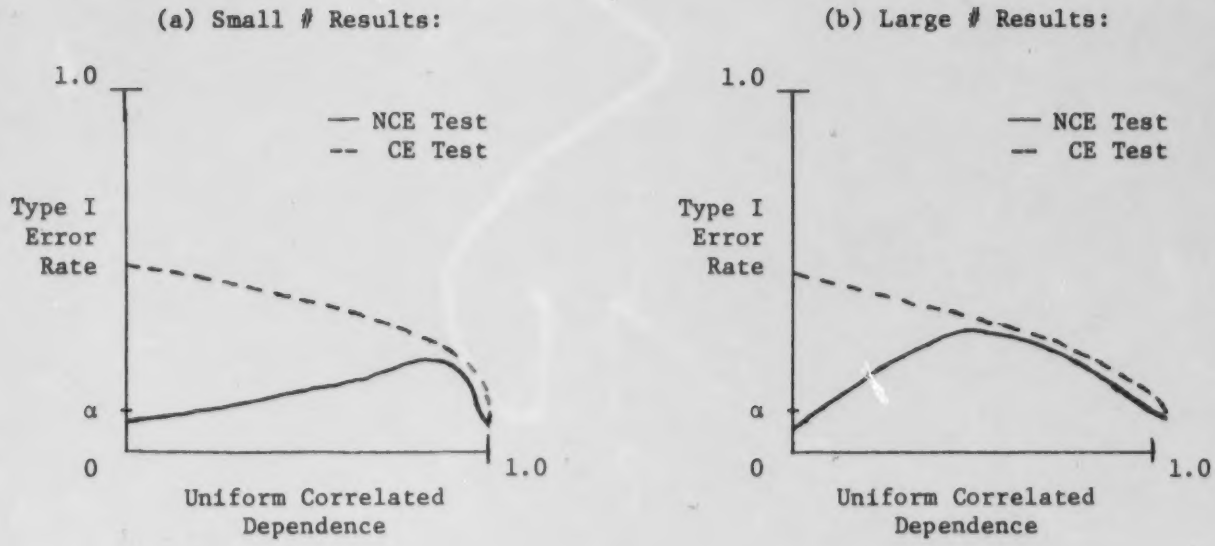


Figure 1. Idealized relationshp between Type I error rate and uniform correlated dependence for (a) a small number of results and for (b) a large number of results.

equals or exceeds the nonchance expectation while the CE test rejects the null hypothesis whenever the number of results exceeds the chance expectation. Table 18 presents the chance expectation and nonchance expectation for varying α (.01, .05, .10 and .25) and for selected set sizes varying from very small (2 results) to very large (10,000 results). In every instance in Table 18, the nonchance expectation exceeds the chance expectation. We will assume that this pattern is generally true for any α and any set size. Thus, the probability of exceeding the chance expectation (a smaller number) is always greater than or equal to the probability of equaling or exceeding the nonchance expectation (a larger number), and the Type I error rate for the CE test will always be greater than or equal to that for the NCE test.

Results displaying uniform correlated dependence will not be found in actual research. In actual research, predictor variable intercorrelations and criterion variable intercorrelations will not all equal the same constant value. Rather, these intercorrelations will vary irregularly with regard to both magnitude and sign. For the present purposes, the type of irregular dependence found in actual research will be termed <u>real data dependence</u>. Example population matrices involving real data dependence are presented in Table 19. The extent or amount of dependence among the predictor-criterion correlations increases as the magnitudes of the predictor-predictor and criterion-criterion correlations in the population matrix increase. For cases involving real data dependence, an estimate of the amount of dependence is provided by the average absolute magnitude of the population values for the predictor-predictor and criterion-criterion correlations. In estimating the amount of dependence, only the absolute magnitudes (unsigned values) of the predictor-predictor and criterion-criterion correlations are considered. Predictor-predictor and criterion-criterion correlations near 0 will produce low dependence while correlations substantially different from 0 (in either

Table 18

Comparison of Chance Expectation (CE)
and Nonchance Expectation (NCE) for Varying Number of Results
and Varying Significance Level (α)

# Results	α = .01		α = .05		α = .10		α = .25	
	CE	NCE	CE	NCE	CE	NCE	CE	NCE
2	.02	2	.10	2	.20	2	.50	2
5	.05	2	.25	2	.50	2	1.25	3
10	.10	2	.50	3	1.00	3	2.50	4
25	.25	3	1.25	4	2.50	5	6.25	9
50	.50	4	2.50	6	5.00	9	12.50	16
100	1.00	5	5.00	10	10.00	15	25.00	29
500	5.00	12	25.00	34	50.00	60	125.00	132
1000	10.00	19	50.00	63	100.00	113	250.00	260
5000	50.00	68	250.00	277	500.00	528	1250.00	1272
10000	100.00	125	500.00	537	1000.00	1040	2500.00	2530

Table 19

Example Population Matrices Involving Real Data Dependence

Matrix for Simulation 30A:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.25	1.0				
P_3	.32	.06	1.0			
C_1				1.0		
C_2				.14	1.0	
C_3				.22	-.18	1.0

Matrix for Simulation 30B:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.26	1.0				
P_3	-.12	.39	1.0			
C_1				1.0		
C_2				.32	1.0	
C_3				.24	.50	1.0

NOTE: Redundant entries above the diagonal have been omitted. Predictor-criterion correlation values are not required for the present purposes and also have been omitted.

a positive or a negative direction) will produce high dependence.

In Table 19, the average absolute magnitude of the predictor-predictor and criterion-criterion correlations is .19 for Simulation 30A and .31 for Simulation 30B. (Note that the 1.0 values in the diagonal of the population matrix are not considered in computing such average magnitudes). It should be cautioned that the average magnitude of the predictor-predictor and criterion-criterion correlations does not provide a full description of the type of dependence found in real data situations. For example, consider the two population matrices in Table 20. The average absolute magnitude of the predictor-predictor and criterion-criterion correlations is .33 for both Matrix A and Matrix B. However, sets of sample results from Matrix A will have a different pattern of interdependences than sets of sample results from Matrix B. It is quite possible that the Type I error rate of a particular test will differ for Matrix A and Matrix B. Thus, the average absolute magnitude of the predictor-predictor and criterion-criterion correlations in the population is only a rough estimate of the amount of dependence among sets of predictor-criterion correlational results.

A series of simulations was undertaken to investigate the influence of real data dependence on Type I error rates for the CE and NCE tests. Table 21 presents the results of simulations performed with $\alpha = .05$ while Table 22 presents the results of simulations performed with $\alpha = .10$. The format for both of these tables should be familiar. The "Dependence" values in both tables are the average absolute

magnitudes of the predictor-predictor and criterion-criterion correlations in the population matrix.[10] In both tables, simulations involving minimally

---

[10] The normally recommended method for averaging correlation coefficients is to transform all correlation values into z values (Fisher's r to z transformation), to average these z values, and then to transform the average z value back into the corresponding correlation value (McNemar, 1962, p. 139). This method was overlooked in arriving at the measure of dependence for real data situations and a simple average of the correlation values was employed. Which averaging method provides a better index of real data dependence is open to question. The simple average is smaller than the average obtained using the r to z transformation. In the present applications, the correlation values to be averaged are small to moderate (correlation with absolute magnitudes exceeding .5 are infrequent) and the difference between the two averages will not be large. Thus, regardless of which type of average is more appropriate, the simple averages employed in this report should provide rough estimates which are satisfactory for the present purposes.

Table 20

Predictor-Predictor and Criterion-Criterion Values
in Two Hypothetical Population Intercorrelation Matrices

Matrix A:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.33	1.0				
P_3	.33	.33	1.0			
C_1				1.0		
C_2				.33	1.0	
C_3				.33	.33	1.0

Matrix B:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	1.0	1.0				
P_3	0	0	1.0			
C_1				1.0		
C_2				1.0	1.0	
C_3				0	0	1.0

NOTE: Redundant entries above the diagonal have been omitted. Predictor-criterion correlation values are not required for the present purposes and have been omitted.

Table 21

Simulations of Sets of Chance-Produced Results
Involving Real Data Dependence and $\alpha = .05$

SIM #	# Results	Dependence*	Type I Error Rate	
			CE Test	NCE Test
17-2A	9	0	.45	.00
30A	9	.19	.34	.00
30B	9	.31	.48	.03
17-4A	25	0	.30	.01
31A	25	.20	.46	.06
31B	25	.26	.41	.08
17-5A	49	0	.41	.08
32A	49	.37	.39	.05
32B	49	.24	.44	.09

NOTE: All simulations involved $\alpha = .05$, 100 samples and 42 Ss per sample. For all simulations except 17-2A, 17-4A and 17-5A, the population correlations for all pairs of predictor variables and for all pairs of criterion variables were correlations obtained in an actual study of classroom education (Borich, Godbout, Peck, Kash and Poynor, 1974). Simulations 17-2A, 17-4A and 17-5A involved minimally dependent results (i.e., all predictor-predictor and criterion-criterion correlations were 0 in the population) and these simulations have been included for comparison purposes.

*The "Dependence" value is the average absolute magnitude of the predictor-predictor and criterion-criterion correlations in the population intercorrelation matrix.

Table 22

Simulations of Sets of Chance-Produced Results
Involving Real Data Dependence and $\alpha = .10$

SIM #	# Results	# Samples	# Ss per Sample	Dependence*	Type I Error Rate	
					CE Test	NCE Test
17-4B	25	100	42	0	.44	.10
13A	25	100	22	.16	.53	.09
13B	25	100	42	.16	.59	.15
13C	25	100	72	.16	.50	.10
17-6B	100	100	42	0	.44	.06
14A	100	100	22	.18	.43	.14
14B	100	100	42	.18	.40	.12
14D	100	500	42	.18	.43	.14
14C	100	100	72	.18	.34	.21
22B	196	100	42	0	.49	.08
21A	196	100	22	.19	.46	.21
21B	196	100	42	.19	.48	.29
21D	196	500	42	.19	.51	.21
21C	196	100	72	.19	.52	.20

NOTE: All simulations involved $\alpha = .10$. For all simulations except 17-4B, 17-6B, and 22B, the population correlations for all pairs of predictor variables and all pairs of criterion variables were correlations obtained in an actual study of classroom education (Borich, Godbout, Peck, Kash and Poynor, 1974). Simulations 17-4B, 17-6B and 22B involved minimally dependent results (i.e., all predictor-predictor and criterion-criterion correlations were 0 in the population) and these simulations have been included for comparison purposes.

*The "Dependence" value is the average absolute magnitude of the predictor-predictor and criterion-criterion correlations in the population intercorrelation matrix.

dependent results (Simulation 17-2A, 17-4A and 17-5A in Table 21 and Simulation 17-4B, 17-6B and 22B in Table 22) have been included for comparison purposes. The remaining simulations in Tables 21 and 22 involved real data dependence, the population predictor-predictor and criterion-criterion correlations being values empirically obtained in an actual educational research investigation (Borich, Godbout, Peck, Kush and Poynor, 1974).[11] The specific population intercorrelation matrices for simulations involving real data dependence are presented in Appendix C.

Consider the CE test Type I error rates in Table 21 ($\alpha = .05$). For simulations involving 25 results, higher CE test Type I error rates were found for .20 and .26 real data dependence than for minimal dependence. However, no corresponding increases in CE test Type I error rate were found for simulations involving 9 results or 49 results. For 9 results and 49 results, the CE test Type I error rates for real data dependence varied around the error rates found for minimally dependent results. Consider the CE test Type I error rates in Table 22 ($\alpha = .10$). For simulations involving 25 results, higher CE test Type I error rates were again found for real data dependence than for minimal dependence. However, for sets of 100 results or

---

[11] A population intercorrelation matrix relevant to real data dependence cannot be constructed simply by randomly assigning values to a population matrix. The values in a correlation matrix are interrelated in complex fashion. In mathematical terms a correlation matrix is a gramian matrix (Guertin & Bailey, 1970). To insure that the population correlation matrices in the simulations involving real data dependence were in fact valid correlation matrices, sets of empirically obtained correlations were used as the predictor-predictor and criterion-criterion values in these population matrices.

196 results, the CE test Type I error rates show little difference for real data dependence and minimal dependence. The discrepancy for 25 results in both Table 21 and Table 22 is quite puzzling, especially given that CE test Type I error rates for 25 results had not previously been found to be higher for uniform correlated dependence than for minimal dependence (see Table 17). Thus, an increase in CE test Type I error rate with increasing dependence does not appear to occur according to an orderly pattern.

Discounting the findings for 25 results, our general conclusion from Tables 21 and 22 is that the CE test Type I error rates for .16 to .31 real data dependence are approximately equal to those for minimal dependence. Had levels of real data dependence above .31 been investigated, we assume that a drop in CE test Type I error rate would have been witnessed for such higher levels of dependence. We base this assumption on the previous simulation findings for uniform correlated dependence--CE test Type I error rate decreased as uniform correlated dependence increased--and on the previous demonstration that CE test Type I error rate must approach α as dependence becomes very high (i.e., approaches 1.0). For the range of real data dependence which was actually studied (.16 to .31), the finding that CE test Type I error rates differ little for minimal dependence and real data dependence is not greatly inconsistent with the previous findings for uniform correlated dependence in Table 17. In Table 17, increases from minimal dependence to .2 uniform dependence usually produced small decreases (.06 or less) in CE test Type I error rate and actually produced a small increase in one instance. Thus, the limited available information is not greatly inconsistent with the assumption that real data dependence and uniform correlated dependence have similar effects on CE test Type I error rate.

The assumption that uniform correlated dependence and real data dependence have similar effects receives stronger support with regard to the

NCE test Type I error rate. Consider the NCE test Type I error rate findings in Table 21. Increases from minimal dependence to real data dependence (in the range of .17 to .31) produced little change for 9 results, a small increase for 25 results, and no consistent change for 49 results. Now consider the NCE test Type I error rate findings in Table 22. Increases from minimal dependence to real data dependence (in the range of .16 to .19) produced little change for 25 results, a moderate increase for 100 results, and a larger increase for 196 results. Our general conclusion from Tables 21 and 22 is that increases from 0 dependence to about .2 or .3 real data dependence tended to produce increases in NCE test Type I error rates with the size of this increase being larger the larger the set size (number of results). This increase and its pattern across different set sizes is highly similar to the previous findings for uniform correlated dependence. Recall that, as uniform correlated dependence increased, there was an initial increase in NCE test Type I error rate and the size of this increase was a positive function of set size.

In conclusion then, the limited evidence tends to support the contention that uniform correlated dependence and real data dependence have similar effects on Type I error rates for the CE and NCE tests. The general relationships between dependence and Type I error rate depicted in Figure 1 are thus assumed to be appropriate for either type of dependence. This ends the discussion of the effects of varying dependence on Type I error rate. We now turn our attention to the effects of varying dependence on statistical power.

Sets of results involving varying dependence--Statistical power. The statistical power for the CE and NCE tests as applied to minimally dependent results has previously been discussed. How does the statistical power of the two tests change as dependence increases above the minimal level? Our initial reaction to this question was to attempt to perform simulation studies of the relationship between dependence and statistical power. Unfortunately, the simulation approach is questionable here.

If dependence is not minimal (i.e., if the predictor-predictor and criterion-criterion correlations in the population are nonzero), then it is often impossible to independently vary the predictor-criterion relationships in the population. Consider the population intercorrelation matrix in Table 23. For convenience in discussion, this matrix has been divided into three submatrices--the P matrix (upper left), the P-C matrix (lower left), and the C matrix (lower right). Note in Table 23 that all predictor-predictor (P matrix) and criterion-criterion (C matrix) correlations are .8. According to present terminology, this population intercorrelation matrix defines sets of predictor-criterion results which have a uniform correlated dependence of .8. Note also in Table 23 that one of the predictor-criterion correlations ($r_{P_1C_1}$) has been assigned a value of 1.0 and all other predictor-criterion correlations (P-C matrix) have been assigned values of 0. At first glance, the matrix in Table 23 may appear to be an appropriate matrix for studying statistical power when there is a uniform correlated dependence of .8 and a single true relationship of 1.0. However, it is impossible for the set of correlations in Table 23 to occur simulataneously. For example, P_1 and C_1 are perfectly correlated and both of these variables should therefore have equal correlations with any third variable. But C_2 is shown to be correlated 0 with P_1 but .8 with C_1. Thus, the matrix in Table 23 is not a possible intercorrelation matrix. The correlations in the matrix are not consistent among themselves (in mathematical terms, the matrix is not a gramian matrix).

Table 23.

Example Population Intercorrelation Matrix

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.8	1.0				
P_3	.8	.8	1.0			
C_1	1.0	0	0	1.0		
C_2	0	0	0	.8	1.0	
C_3	0	0	0	.8	.8	1.0

Note: Redundant entries above the diagonal have been omitted.

This problem of inconsistent correlations would arise regardless of the placement of the single true relationship in the P-C matrix. That is, similar inconsistencies would arise if one predictor-criterion value was 1.0 and all others 0, regardless of whether the 1.0 value was for P_2 and C_1, P_3 and C_1, P_1 and C_2, or any other predictor-criterion pair.

Given a set of predictor-criterion correlations with .8 uniform correlated dependence, it is impossible for one of these correlations to be 1.0 and the other correlations to be 0. If one of the correlations is assumed to be 1.0, then other correlations in the set are necessarily greater than 0. This state of affairs is quite reasonable. If the correlations in a set are interdependent, then the value of any particular correlation should naturally influence the possible range of values for the other correlations.

Let us further examine the manner in which the value of one correlation in a set of dependent results is influenced by the values of the other correlations in that set of results. In Appendix D, we present McNemar's (1962) method for calculating the limits placed upon a correlation between two variables by the values of the correlations of each of these two variables with a third variable. McNemar's method is derived from the fact that partial correlation coefficients involving the three variables must be equal to or greater than -1.0 and less than or equal to +1.0 (see Appendix D). Also in Appendix D, we present an extension of McNemar's logic to cases involving the limits placed upon a correlation between two variables by the values of the correlations of each of these two variables with not only a third variable but also a fourth and a fifth variable. These methods from Appendix D were used to calculate the limits on the values of correlation coefficients in the correlation matrices presented in Table 24.

Consider the matrices in Table 24. All three matrices involve four predictor variables and a single criterion variable. In all three matrices, all predictor variables are intercorrelated at +.8 and therefore the predictor-criterion

Table 24.

Matrix A:

	P_1	P_2	P_3	P_4	C_1
P_1	1.0				
P_2	.8	1.0			
P_3	.8	.8	1.0		
P_4	.8	.8	.8	1.0	
C_1	.8	?			1.0

Matrix B:

	P_1	P_2	P_3	P_4	C_1
P_1	1.0				
P_2	.8	1.0			
P_3	.8	.8	1.0		
P_4	.8	.8	.8	1.0	
C_1	.8	?	.8		1.0

Matrix C:

	P_1	P_2	P_3	P_4	C_1
P_1	1.0				
P_2	.8	1.0			
P_3	.8	.8	1.0		
P_4	.8	.8	.8	1.0	
C_1	.8	?	.8	.8	1.0

Note: Redundant entries above the diagonal have been omitted.

correlations in each matrix demonstrate uniform correlated dependence of +.8. In matrix A, a single true relationship of +.8 has been assumed (for $r_{P_1C_1}$); in matrix B, two true relationships of +.8 have been assumed (for $r_{P_1C_1}$ and $r_{P_3C_1}$); and in matrix C, three true relationships of +.8 have been assumed (for $r_{P_1C_1}$, $r_{P_3C_1}$ and $r_{P_4C_1}$). In each of these three cases, we are interested in the possible values of $r_{P_2C_1}$ given the assumed true relationships. The methods in Appendix D allow calculation of the possible range of values for $r_{P_2C_1}$ in each of these three matrices.

Consider matrix A in Table 24. What is the possible range of values for the P_2C_1 correlation in this matrix? Expressions [1] and [2] in Appendix D allow calculation of the upper and lower bounds of $r_{P_2C_1}$ as a function of the assumed values for $r_{P_1P_2}$ and $r_{P_1C_1}$. Performing such calculations, we obtain +.28 as the lower bound for $r_{P_2C_1}$ and +1.0 as the upper bound for this correlation. Note that a 0 value for $r_{P_2C_1}$ is impossible, this correlation necessarily being +.28 or greater.

Consider matrix B in Table 24. What is the possible range of values for $r_{P_2C_1}$ in this matrix? Expressions [7] and [8] in Appendix D allow calculation of the upper and lower bounds for $r_{P_2C_1}$ as a function of the assumed values for $r_{P_1P_2}$, $r_{P_1C_1}$, $r_{P_2P_3}$ and $r_{P_3C_1}$. Performing such calculations, we obtain +.42 as the lower bound for $r_{P_2C_1}$ and +1.0 as the upper bound for this correlation.

Finally, consider matrix C in Table 24. What is the possible range of values for $r_{P_2C_1}$ in this Matrix? Expression [10] in Appendix D allows calculation of the upper and lower bounds of $r_{P_2C_1}$ as a function of the assumed values for $r_{P_1P_2}$, $r_{P_1C_1}$, $r_{P_2P_3}$, $r_{P_3C_1}$, $r_{P_2P_4}$ and $r_{P_4C_1}$. The calculated range of possible values for $r_{P_2C_1}$ is +.46 to +1.0.

These three examples illustrate that, when dependence is high (.8), then it is impossible to independently vary the population relationships. If one or more strong true relationships of +.8 are assumed, then other predictor-criterion correlations cannot be 0. Furthermore, the lower bound for a predictor-

criterion correlation systematically increases as the number of assumed +.8 relationships increases.

The methods in Appendix D can be applied to matrices involving lower dependence. For example, consider the case in which we have a uniform correlated dependence of .6. In this case, the lower bound for a predictor-criterion correlation is 0 if one of the other predictor-criterion correlations is assumed to be +.8 and this lower bound will increase above 0 if two or more of the other predictor-criterion correlations are assumed to be +.8. Thus, when dependence is .6, it is impossible to independently establish 2 or more strong true relationships of +.8 and allow the remaining predictor-criterion correlations to all be 0. Such examples illustrate the infeasibility of performing simulation studies of statistical power under conditions of high dependence. Population matrices involving a fixed level of high dependence (e.g., .6 or .8) and a predetermined number of true relationships (with all other predictor-criterion correlations being 0) may not be valid correlation matrices.

It is difficult to perform simulation studies of statistical power when there is a high level of dependence. Does such a difficulty also occur for low dependence? We could try to answer this question with the methods in Appendix D for calculating the bounds of a correlation coefficient in a correlation matrix. However, such methods would not lead to the correct conclusion. The methods in Appendix D are of limited usefulness especially with regard to assessing the validity of low dependence matrices. A better method is provided by attempted application of the present simulation procedures and this method will be used for low dependence matrices. In fact, attempted application of the simulation methods could have been used to demonstrate the difficulties encountered in performing simulation studies of statistical power when dependence is high. However, we have chosen to demonstrate the difficulties for high dependence with the methods in Appendix D, such methods providing some insight into the complex interrelationships among the correlations within a correlation matrix.

The methods in Appendix D are based upon the fact that a partial correlation coefficient cannot be less than -1 or greater than +1. More specifically, the value of r_{12} must be such that partial correlation coefficients of the form $r_{12.3}$, $r_{12.34}$ and $r_{12.345}$ are not less than -1 or greater than +1.[12] Such restrictions on the value of r_{12} are not the only restrictions on the value of that correlation. Given a five variable intercorrelation matrix, the value of r_{12} is restricted by the following facts: (1) the value of any multiple correlation involving the 5 variables or a subset of the 5 variables must fall in the range 0 to 1; (2) the value of any semipartial correlation coefficient involving the 5 variables or a subset of the 5 variables must fall in the range of -1 to +1, and; (3) the value of any partial correlation coefficient (not only of the form $r_{12.34}$ but also of the form $r_{34.12}$) involving the 5 variables or a subset of the 5 variables must fall in the range -1 to +1. Thus, the methods in Appendix D provide only a partial check on the consistency of a set of correlation coefficients. These methods consider only a subset of the restrictions on the values in a correlation matrix.

A more thorough method of determining whether a set of correlation values are consistent is provided by attempted application of the simulation procedures presented in this report. The matrix of correlation values in question is submitted to our simulation program as a population intercorrelation matrix. If the matrix is a valid intercorrelation matrix, then correlation values obtained in the simulated samples should be close to the population values

---

[12]Note that the methods in Appendix D establish bounds for r_{12} on the basis of partial correlation coefficients involving the relationship between Variable 1 and Variable 2 with additional variables serving as control variables--i.e., partial correlation coefficients such as $r_{12..34}$. These methods do not apply to the limits placed on r_{12} by partial correlation coefficients involving Variable 1 and/or Variable 2 as control variables--i.e., partial correlation coefficients such as $r_{34.12}$.

submitted to the program.[13] If the matrix is not a valid correlation matrix, then the population matrix will not be reproduced in the simulated samples.[14]

The present simulation techniques were used to determine the feasibility of studying statistical power when dependence is low (e.g., .2). Table 25 presents relevant simulation results. The matrix of values in Table 25 presents the average predictor-criterion correlations obtained over the 100 samples in Simulation 28-B.[15] The population matrix for Simulation 28-B involved uniform correlated dependence (i.e., all off-diagonal predictor-predictor and criterion-criterion correlations were .2) and 5 true predictor-criterion relationships of +.7 with all other predictor-criterion relationships being 0. The 5 predictor-criterion correlations with +.7 population values have been underlined in the matrix in Table 25. The average simulated correlations in Table 25 indicate a failure to reproduce the population matrix. Many of the correlations with 0 population values have average simulated values of +.1 or +.2. Thus, the population matrix for Simulation 28-B is not a valid correlation matrix. The results in Table 25 indicate the impossibility of independently manipulating population relationships when dependence is .2.

Consider another example involving low dependence. Table 26 presents

---

[13] In the appendix concerning the validity of the present simulation procedures (Appendix B), we have seen that simulated sample values reproduce a valid population intercorrelation matrix within the limits of sampling error.

[14] Our simulation procedures are based upon the diagonal or triangular factorization technique. Application of such factor analytic methods to a matrix which is not an intercorrelation matrix (a nongramian matrix) will result in a nonsensical factor structure which will not allow reproduction of the original matrix.

[15] These average predictor-criterion correlations represent simple averages of the sample values rather than the more appropriate method based upon the Fisher r to z transformation (McNemar, 1962). The Fisher r to z method was overlooked when the present simulation program was written. The simple average of the sample values will approximately equal a population value which is close to 0 but may differ greatly from a population value which is not close to 0. When the population value is quite high, then the simple average across samples will be closer to 0 than the population value. For example, an average sample value of +.5 or +.6 will be obtained when the population value is +.7.

Table 25.

Simulation 28-B: Mean Predictor-Criterion Correlations Obtained Across 100 Samples with 42 Subjects per Sample

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
C_1	.5	.1	.6	.2	.2	.2	.2	.2	.1	.2
C_2	.0	.5	.0	.6	.2	.2	.2	.2	.2	.2
C_3	.0	.0	.0	.0	.0	.0	.5	.1	.1	.1
C_4	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_5	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_6	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_7	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_8	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_9	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_{10}	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0

Table 26.

Simulation 34-A: Population Matrix and Mean Simulated Values Across 100 Samples with 42 Subjects per sample

Population Matrix:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.25	1.0				
P_3	.32	.06	1.0			
C_1	.70	0	0	1.0		
C_2	0	0	0	.14	1.0	
C_3	0	0	0	-.22	-.18	1.0

Mean Simulated Values:

	P_1	P_2	P_3	C_1	C_2	C_3
P_1	1.0					
P_2	.24	1.0				
P_3	.32	.15	1.0			
C_1	.57	.13	.19	1.0		
C_2	.01	.00	-.01	.11	1.0	
C_3	-.01	-.01	-.03	-.19	-.21	1.0

Note: Redundant entries above the diagonals have been omitted.

information relevant to Simulation 34-A. The upper matrix in this table is the population matrix for that simulation. Note in the population matrix that the predictor-criterion results display the type of irregular dependence found with real data--the predictor-predictor and criterion-criterion values in this matrix are actual correlation values obtained by Borich, Godbout, Peck, Kash and Poynor (1974). The dependence among sets of predictor-criterion results defined by this population matrix is low, the average absolute magnitude of the off-diagonal predictor-predictor and criterion-criterion correlations in the population matrix being 0.20. Also note in the population matrix that a single true predictor-criterion relationship of +.7 has been assumed with all other predictor-criterion relationships being 0.

The lower matrix in Table 26 presents the mean simulated correlation values over 100 samples. Note that the predictor-predictor and criterion-criterion average simulated values in the lower matrix are quite close to the corresponding population values in the upper matrix. Such close correspondence between simulated and population values did not occur for all of the predictor-criterion correlations. For two of the predictor-criterion correlation ($r_{P_2C_1}$ and $r_{P_3C_1}$), the population values were 0 but the average simulated values were .13 and .19, values well above 0. Thus, the population matrix for Simulation 34-A is not a valid correlation matrix. The results for this simulation further confirm the difficulty of independently manipulating population relationships when dependence is low.

In summary then, it is not possible to independently manipulate the population relationships when there is correlated dependence, whether such dependence is low (e.g., .2) or high (e.g., .6 or .8). This fact makes it difficult to use the present simulation techniques to systematically study statistical power under conditions of varying dependence.

It should be mentioned that, when dependence is minimal (all nondiagonal predictor-predictor and criterion-criterion values in the population

intercorrelation matrix are 0), then it is often possible to independently manipulate population relationships. For example, consider the matrix in Table 27. This matrix presents the average predictor-criterion correlations obtained across the 100 samples in Simulation 28-A. The population matrix for this simulation involved minimal dependence, 5 true predictor-criterion relationships of +.7 (corresponding to the 5 underlined values in Table 27), and 95 predictor-criterion correlations of 0. The average simulated values in Table 27 correspond closely to this population matrix and we assume that this population matrix is a valid intercorrelation matrix. This example simulation illustrates the possibility of independently manipulating population relationships when dependence is minimal.

In addressing the possible use of simulations to study statistical power, we have limited our attention to strong relationships in the population (e.g., population relationships of .7 or .8). The difficulty in studying power when dependence is not minimal and when the population relationships are strong may diminish if we shift our attention to weaker population relationships. It may be possible to form a valid population intercorrelation matrix with correlated dependence, a specified number of weak predictor-criterion relationships, and all other predictor-criterion relationships equal to 0. Unfortunately, in the present report we have been preoccupied with strong population relationships and the possibility of studying weak population relationships has occurred to us only as an afterthought.

The main purpose of the present section of this report is to examine the relationship between statistical power and dependence. However, we have yet to present any information about this relationship. Up to this point, we have been concerned with the possibility of performing simulation studies of this relationship, and we have encountered a problem--the impossibility of independently manipulating the population relationships for a set of results involving correlated dependence. While posing difficulties for a simulation

Table 27.

Simulation 28-A: Mean Simulated Correlations
Across 100 Samples with 42 Subjects per Sample

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
C_1	.5	10	.5	.0	.0	.0	.0	.0	.0	.0
C_2	.0	.5	.0	.5	.0	.0	.0	.0	.0	.0
C_3	.0	.0	.0	.0	.0	.0	.6	.0	.0	.0
C_4	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_5	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_6	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_7	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_8	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_9	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
C_{10}	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0

approach, the very nature of this problem suggests the following argument concerning the relationship between statistical power and dependence.

When results are minimally dependent, then nonzero population relationships can be independently established. A strong population relationship assumed for one predictor-criterion pair does not necessitate that other predictor-criterion pairs have nonzero population relationships. The existence of a single strong population relationship should cause only one of the sample predictor-criterion results to be significant with a greater than chance frequency. In constrast, when results display correlated dependence (dependence greater than minimal dependence), then such independent manipulation of the population relationships is not possible. A strong population relationship established for one predictor-criterion pair will necessitate that other predictor-criterion pairs also have nonzero population relationships. The existence of a single strong population relationship should cause several of the sample predictor-criterion results to be significant with a greater than chance probability. When dependence is at a maximum (all predictor-predictor and criterion-criterion population values equal to 1.0), then a strong population relationship established for one predictor-criterion pair necessitates that equally strong population relationships also exist for all the other predictor-criterion pairs.[16] The existence of a single strong population relationship should cause all of the sample predictor-criterion results to be significant with a greater than chance probability.

When dependence is minimal, a true population relationship will be reflected in only <u>one</u> sample result; when dependence is greater than minimal, a true population relationship can be reflected in <u>several</u> sample results; and when

[16] When dependence is 1.0, then all predictor-criterion correlations in the population must be equal.

dependence is at a maximum, a true population relationship will be reflected in all sample results. Given a single true population relationship, the number of significant sample results should tend to be greater the higher the dependence. It therefore follows that the probability of detecting true population relationships with the present tests for analyzing sets of results (the CE test and the NCE test) should increase as dependence increases. That is, the statistical power of both the CE test and the NCE test increases with increasing dependence.

In the previous section on statistical power for minimally dependent results, high power was defined as a probability greater than 0.80 of detecting a proportion of strong population relationships equal to α or less. Such high power was adopted as a minimum standard for the recommended use of a test. The positive relationship between statistical power and dependence implies that a test which achieves high power in application to a set of minimally dependent results will also achieve high power when applied to a comparable set of results characterized by dependence above the minimal level.

No attempt will be made to explore the relationship between statistical power and dependence in finer detail, and the present section on statistical power for results involving varying dependence will now be concluded. The following section represents an attempt to formulate use recommendations for the CE test and NCE test in application to sets of results displaying any type of dependence. Formulation of such use recommendations requires simultaneous consideration of both the Type I error rate and the statistical power under varying conditions of dependence. Thus, the following section is an attempt to integrate the main finding of the present section on statistical power for varying dependence results with the major findings of the preceding section on Type I error rate for varying dependence results.

Sets of results involving varying dependence--Summary and conclusions. It is our contention that the use recommendations presented in Table 15 for minimally dependent results are appropriate for results displaying any type of dependence. Our primary consideration in recommending a test for use in a given application is that it achieves a high level of power in that application. In the immediately preceding section, we have seen that statistical power increases as dependence increases. If a test achieves high power for a set of minimally dependent results, then it will also achieve high power for comparable sets of results involving greater than minimal dependence. Thus, a test which has been recommended for minimally dependent results can also be recommended for comparable sets of results involving any type of dependence.

While a high level of statistical power is the primary concern in recommending the use of a test, a secondary concern is a low Type I error rate. If both the CE test and the NCE test achieve high power for a given application, then the test with the lower Type I error rate is to be recommended. Preceding sections of this report have indicated that (1) for minimally dependent results, the Type I error rate for the NCE test is less than that for the CE test and (2) for results displaying greater than minimal dependence, the Type I error rate for the NCE test is less than or equal to that for the CE test. Thus, considering Type I error rate, a test which has been recommended for minimally dependent results can also be recommended for comparable sets of results involving any type of dependence.

In general then, the use recommendations for minimally dependent sets of results (Table 15) are appropriate for sets of results involving any type of dependence. The use recommendations from Table 15 have been reproduced in Table 28 as general use recommendations. Note that no test is recommended for small numbers of results (less than 50 results for $\alpha = .05$ and less than

Table 28.

Analysis Methods Recommended as Appropriate for
Sets of Results Involving any Type of Dependence

# Results	Test Recommended for $\alpha = .05$	Test Recommended for $\alpha = .10$
less than 25	no test	no test
25	no test	CE test
50	CE test	CE test
100	CE test	NCE test
200 or more	NCE test	NCE test

25 results for $\alpha = .10$), the CE test is recommended for a moderate number of results (50 or more but less than 200 results for $\alpha = .05$ and 25 or more but less than 100 results for $\alpha = .10$), and the NCE test is recommended for a relatively large number of results (200 or more results for $\alpha = .05$ and 100 or more results for $\alpha = .10$). These use recommendations in Table 28 are extremely important. They represent the climax of our study of the possibility of analyzing sets of results.

A more detailed discussion of the logic used in formulating the present use recommendations can be found in a previous section of this report entitled "Minimally Dependent Sets of Results--Summary and Conclusions." Included in that section are arguments concerning the priority given statistical power in recommending a test as well as arguments for the reasonableness of our adopted standard of high power.

The validity of the recommendations in Table 28 rests largely on the validity of the simulation results which have been presented. A few words of caution are in order concerning the simulation results. First, the range of simulation studies undertaken has been limited, and we have attempted to identify general trends or relationships on the basis of the limited simulation results. For example, none of the present simulations involved more than 196 results[17] or significance levels other than .05 or .10. While we have no reason to suspect that different general trends or relationships would be found for larger sets of results or different significance levels, we have not tested this possibility. A second caution involves the validity of the simulation procedures. Appendix B presents information and discussion relevant to the validity of the simulation procedures. Most of the evidence in Appendix

[17]The present simulations were limited to relatively small numbers of results (i.e., 196 or fewer results) largely because of the high computer time cost in simulating larger sets of results.

B suggests that the simulation procedures produced sample predictor-criterion results which differed from population values within the range of sampling error. However, Appendix B also presents some evidence for irregularities in the simulation results. While the general conclusion in Appendix B is that the simulation results are valid for the present purposes, the reader should be aware that some irregularities do exist.

This concludes discussion of the feasibility of analyzing sets of results. Our conclusion is that both the CE and the NCE tests are useful in performing overall evaluations of certain sets of results. We now turn our attention to the range of applications of the two tests and then to example applications.

Analysis of Sets of Results--Range of Applications

The CE and NCE tests provide methods useful in multivariate research. However, these two tests are unnecessary in multivariate studies which have large sample size relative to number of variables analyzed. Such multivariate studies are appropriately analyzed with the standard multivariate tests for assessment of overall relationship (e.g., canonical correlation or multivariate analysis of variance). In contrast, such standard multivariate techniques are not applicable in studies involving small sample sizes and many variables. This latter type of study, typically an exploratory study, is where the CE and NCE tests find usefulness. In exploratory studies, the CE and NCE tests provide the only methods for determining if the entire set of results is probable on the basis of chance or not. In the present report, our main interest is the application of these two tests to exploratory studies of classroom education, but, these methods may also be useful in exploratory research in other fields.

Analysis of types of results other than predictor-criterion correlations. Up to this point, we have considered application of the CE and NCE tests only

to sets of predictor-criterion correlations. However, these tests are applicable to virtually any set of inferential statistical results. For example, these tests could be applied to sets of partial correlation coefficients, sets of semipartial correlation coefficients, sets of *t*-test results, sets of analysis of variance results, sets of analysis of covariance results, or sets of results derived by using more complex multiple regression analysis techniques. To apply either test to any set of results, the only information necessary is (1) the total number of results in that set, (2) the number of significant results in that set, and (3) the level of significance--i.e., α.

The general use recommendations for the CE and NCE tests in Table 28 are based upon the study of sets of simple predictor-criterion correlations. Such predictor-criterion correlations were chosen for study because of the ease of defining and manipulating the level of dependence for such results, the level of dependence within a set of predictor-criterion correlations being a function of the population values of the predictor-predictor and criterion-criterion correlations. Defining and manipulating the level of dependence of, say, analysis of variance or analysis of covariance results would be more complicated. However, we find no reason to suspect that the trend of the present Type I error rate and statistical power findings is not generalizable to any type of inferential statistical results. In conclusion then, we suggest that the general use recommendations in Table 28 are appropriate not only for predictor-criterion correlations but also for any other type of inferential statistical results.

Analysis of subsets of results. Consider a hypothetical exploratory study of classroom education involving 1000 predictor-criterion correlations, a small sample size[18] (e.g., 40 classrooms with classroom serving as the unit of analysis), and a significance level (α) of .05. Since the NCE test is recommended for use

[18] A small sample size has been assumed in this hypothetical study in order to rule out possible application of standard multivariate tests (e.g., multivariate analysis of variance or canonical correlation).

with 200 or more results when $\alpha = .05$ (see Table 28), this test can be used to analyze the set of 1000 results in this hypothetical study. Suppose that 100 of the 1000 results are significant. Since the nonchance expectation for 1000 results with $\alpha = .05$ is only 63 (as given in Appendix A), it is extremely improbable that all 100 significant results could have occurred by chance. The obtained number of significant results exceeds the nonchance expectation and we therefore reject the null hypothesis that the entire set of results is chance-produced. It is quite likely that many of the significant results reflect valid relationships.

Given that the overall significance for the entire set of 1000 results implies the existence of valid relationships, it would be valuable to further specify the source of the valid relationships. Suppose that 500 of the 1000 predictor-criterion correlations in this hypothetical study involve measures of reading achievement as criteria while the remaining 500 correlations involve measures of mathematics achievement as criteria. Do valid relationships occur only for reading achievement, only for mathematics achievement or for both reading and mathematics achievement? To answer this question, the NCE test could be separately applied to the subset of reading achievement results and to the subset of mathematics achievement results. Suppose that in our hypothetical study, the 100 significant results are found to be divided 70 and 30 between reading achievement and mathematics achievement, respectively. In Appendix A, the nonchance expectation can be seen to be 34 for 500 results with $\alpha = .05$. Therefore, the null hypothesis that the entire subset of results is chance-produced would be rejected for reading achievement but not for mathematics achievement. The resulting conclusion would be that the valid relationships tended to occur for reading achievement but not for mathematics achievement.

Given the lack of significance for the mathematics achievement subset,

it would seem pointless to examine smaller subsets within mathematics achievement. In contrast, the significance for the reading achievement subset invites further examination of that subset. The reading achievement results could be divided into smaller subsets and an overall analysis of each of these smaller subsets could then be performed. Given overall significance for one of these smaller subsets of results, that subset could be further divided into yet smaller subsets which could then be analyzed with the CE or NCE test. Such analysis of smaller and smaller subsets could continue until the subsets became so small that the CE test and NCE test would no longer be recommended. (In our hypothetical study with $\alpha = .05$, neither test is recommended for application to less than 50 results).

The interpretation of the results of this hypothetical study would proceed in the following fashion. The most important result is clearly the overall significance for the entire set of 1000 results. This significant overall result supports the hypothesis that there is better than chance prediction of criterion performance in the study. Had this overall result been non-significant, then interpretation of the results of the study would have stopped at this point. However, given the overall significance, the next step in interpreting the results would involve consideration of subsets of the results. Given the significance for the subset of 500 reading achievement results but the lack of significance for the subset of 500 mathematics achievement results, it can be concluded that the better than chance prediction of criterion performance is limited to reading achievement. At this point, it is no longer necessary to consider the mathematics achievement results, any significant predictor-criterion correlations involving mathematics achievement measures being interpreted as chance-produced. The next step would be consideration of subsets of the reading achievement results. Given a nonsignificant subset of the reading achievement results, any significant

correlations in this subset would be interpreted as being chance-produced. Given a significant subset of the reading achievement results, the significant predictor-criterion correlations in that subset are taken as indicating possible valid relationships. This process would continue with smaller and smaller subsets until the smallest subsets which were analyzed have been considered. Note that at the end of this process, the significant predictor-criterion results in the study will have been partitioned into two groups. The first group involves all significant correlations which have been included in a nonsignificant subset at any step of the process, such significant correlations having been identified as chance-produced. The second group involves all significant correlations which have never been included in a nonsignificant subset, such significant correlations having been identified as most likely to reflect valid relationships.

The example of this hypothetical study illustrates a useful and important scheme for analyzing and interpreting sets of results. The logic underlying this scheme is similar to that found with regard to other statistical procedures. For example, in studies allowing the use of standard multivariate tests (e.g., multivariate analysis of variance or canonical correlation), it is often recommended (e.g., Cooley & Lohnes, 1971; Bock, 1975; Kerlinger & Pedhazur, 1973) that an overall multivariate test initially be performed for the entire variable set. If this overall test is nonsignificant, then the appropriate conclusion is that the entire data set is the product of chance. Only if the overall test is significant, do these authors recommend consideration of the results for individual variables or subsets of the variables. Similarly, in studies employing analysis of variance, it is often recommended (e.g., Edwards, 1968; Glass & Stanley, 1970) that the finding of a significant overall treatment effect be followed by multiple comparison tests. Such multiple

comparison tests involve analysis of subsets of the treatment means and allow identification of the specific mean differences producing the significant overall treatment effect. Note that the use of such multiple comparison tests is not recommended when the overall treatment effect fails to achieve statistical significance. Thus, the present scheme for analyzing sets of results--i.e., an overall evaluation of the entire set of results followed by analysis of subsets of the results--is certainly in line with standard practice.

A few suggestions are in order concerning the formation of subsets of results. First, subsets of results should be formed on a conceptual basis--i.e., the results in a subset should have some meaningful conceptual similarity. Examples of reasonable subsets of results are (1) the subset of results involving reading achievement measures as criteria, (2) the subset of results involving vocabulary scores as criteria, (3) the subset of results involving measures of a particular type of teacher behavior (e.g., criticism) as predictors, (4) the subset of results involving measures of teacher personality as predictors, and (5) the subset of results involving teacher personality measures as predictors and vocabulary scores as criteria. Such subsets of results allow meaningful interpretation. In contrast, subsets defined on a more arbitrary basis--e.g., the subset of results involving odd numbered variables or the subset of results involving the first ten variables--will lead to interpretation difficulties.

Another suggestion concerning formation of subsets of results is that such subsets be mutually exclusive rather than overlapping. Two mutually exclusive subsets contain no members in common while two overlapping subsets contain common members. Consider the case in which there are two overlapping subsets of results, with one subset displaying overall significance but the other subset displaying overall nonsignificance. Individual significant

results within the subset displaying overall significance would normally be interpreted as reflecting possible valid relationships, and individual significant results within the nonsignificant subset would normally be interpreted as products of chance. Individual significant results common to both of the overlapping subsets are difficult to interpret. The evidence is ambiguous concerning whether such significant results are chance-produced or whether they reflect possible valid relationships. Note that such interpretation problems will not occur for mutually exclusive subsets of results. The ease of interpretation for mutually exclusive subsets is the basis for our favoring such subsets.

The present scheme for analyzing a set of results followed by analysis of successively smaller subsets is quite important. When applied to chance-produced sets of results, this scheme will often allow the researcher to avoid the waste of time entailed in interpreting the spuriously significant individual results. When applied to sets of results involving a number of true population relationships, this scheme allows the investigator to systematically identify individual significant results which are likely to be chance-produced and to isolate those significant results most likely to be reflective of valid population relationships.

Analysis of replicated results. Consider a large-scale exploratory study which has been repeated in two separate replications. The replications may have been performed, for example, in two separate school years or in two separate school districts. In such a case, it is often important to know whether significant overall prediction of criterion performance occurs for the combined set of results from the two replications. This research question can be addressed by applying the CE test or the NCE test to the combined set of results.

Such a combined test may offer a statistical power advantage over simply

performing separate tests for the two replications. Recall that the statistical power of the CE and NCE tests increases as the number of results in a set increases. Since the combined set of results involves twice as many results as the individual set for either replication, a test of the combined results will have greater power than separate tests for the two replications. A smaller percentage of valid relationships will be detected in the combined test. Note that such combined tests are not limited to two replications. Combined tests can also be performed for three, four, or more replications.

Again consider a large-scale exploratory investigation which has been repeated in two separate replications. In addition to questions concerning combined sets of results, it may also be of value to address questions concerning the set of joint results for the two replications. A joint result is composed of two individual results, one result being from the first replication and the second result being the corresponding result from the second replication. Of special interest here are joint results involving significance in the same direction for both replications. Is it probable that the number of joint results, involving significance in the same direction for both replications, could have occurred by chance? This research question can be addressed by applying the CE or NCE test to the set of joint results. The information necessary for the application of either test is (1) the total number of joint results, (2) the number of joint results involving significance in the same direction for both replications, and (3) the chance probability for a joint result involving significance in the same direction for both replications.

It is beyond the scope of the present report to consider which research question (combined results vs. joint results) is more appropriate in the case of a replicated investigation. This ends the discussion of the range of applications of the CE and NCE tests. We now turn our attention to example applications of these tests to actual studies of classroom education.

Analysis of Sets of Results--Example Applications

Five examples of analyzing sets of classroom education research results will be considered. Three of these examples--the Soar and Soar (1973) study; the Borich, Godbout, Peck, Kash and Poynor (1974) study; and the Calkins, Godbout, Poynor and Kugle (1976) study--are cases in which the original investigators analyzed sets of results and reported the outcomes. For these three examples, we will describe the methods used and evaluate these methods in light of the information in the present report. Sets of results were not analyzed by the investigators in the other two examples--the Good and Grouws (1975) study and the Brophy and Evertson (1974a; 1974b) study. In these latter two examples, we have analyzed sets of the results reported by the original investigators. Of major interest here is the extent to which analyses of sets of results suggest different conclusions than those presented by the original investigators.

The Soar and Soar (1973) study. The Soar and Soar investigation represents one of the earliest reported uses of methods for analyzing sets of results. These authors were quite aware of the problem of spurious significance and the need for analyzing sets of results, as evidenced by the following quotation (Soar and Soar, 1973, p. 44):

> The number of tests run of course raises the problem that significant results would be found by chance. This problem is well known. But knowledge of the educational domain is so limited, and theory so little help in focusing on specific hypotheses, that the wiser course seemed to be a broad-ranging screening process, recognizing that not all apparently significant results would have meaning. But instead of looking for single significant results, agreement across numbers of relationships was examined and discussed.

The purpose of the Soar and Soar investigation was to study pupil performance as a function of observed classroom behavior. This investigation involved two fifth grade samples (59 classrooms and 22 classrooms,respectively)

and also a second grade sample (23 classrooms). The main results of the investigation concerned the fifth grade sample of 59 classrooms and we will limit our attention to this sample. Seventeen pupil performance measures were studied including measures of subject matter competence, motivation, creativety, intelligence, anxiety and dependence-proneness. Pupil performance was assessed on three separate occasions--(1) at the beginning of the academic year, (2) at the end of the academic year, and (3) at the end of the following summer. Two residual gain scores were computed for each pupil performance measure--(1) gain over the academic year and (2) gain over the summer months. The resulting 34 gain measures (gain in 17 pupil performance measures over two different time periods) served as criterion variables. The predictor variables were 11 measures of observed classroom behavior arrived at on the basis of a series of factor analyses of several hundred variables from five different classroom observation systems.

To assess linear predictor-criterion relationships, all possible pairs of predictor-criterion correlations involving the 11 predictor variables and the 34 criterion variables were computed (a total of 374 correlations). In addition, nonlinear relationships for each predictor variables and interactions involving pairs of predictor variables were also assessed with regard to each criterion variable. Of interest was the extent to which a nonlinear or interaction variable allowed prediction of criterion performance over and above that achieved with the original 11 linear predictor variables. Multiple regression analysis techniques were used to assess the significance of such prediction for each of 28 nonlinear and interaction variables paired with each of the 34 criterion variables (a total of 952 significance tests for nonlinear relationships and interactions).

A total of 1326 significance tests were reported by Soar and Soar (374 tests for linear relationships and 952 tests for nonlinear relationships and

interactions). All of these analyses involved $\alpha = .05$ with the classroom as the unit of analysis. The authors failed to perform an overall analysis of the entire set of 1326 results and we will now perform such an analysis. With 1326 results and $\alpha = .05$, the NCE test is recommended (see Table 28) and the nonchance expectation is 81 (see Appendix A). This nonchance expectation is far exceeded by the 233 significant results reported by Soar and Soar and the obvious conclusion is that the majority of these significant results must reflect valid population relationships.

Given significance for the entire set of results, the next logical step would be to analyze subsets of results. Soar and Soar report the analyses of four subsets of results--Subset 1, the 187 linear results for gain over the academic year; Subset 2, the 187 linear results for gain over the summer; Subset 3, the 476 nonlinear and interaction results for gain over the academic year; and Subset 4, the 476 nonlinear and interaction results for gain over the summer. To analyze these subsets of results, Soar and Soar used a method corresponding to what we have labeled the CE test in the present report.[19] Their findings were overall nonsignificance for Subset 2 (8 obtained significant results with a chance expectation of 9.35) but overall significance for the other three subsets.

Soar and Soar did not further divide the three significant subsets in an attempt to illiminate additional chance-produced significant results. They simply accepted all significant results in Subsets 1, 3 and 4 as valid and proceeded to interpret these results. It is possible that analysis of smaller

---

[19] According to the recommendations in Table 28 for $\alpha = .05$, the CE test should be applied to 50 or more but less than 200 results and the NCE test should be applied to 200 or more results. We would, therefore, recommend the CE test for Subsets 1 and 2 but the NCE test for Subsets 3 and 4. Soar and Soar applied the CE test to all four subsets. However, for the case in question, the NCE and CE tests will both lead to the same conclusions for Subset 3 as well as for Subset 4.

subsets of results would have further isolated the significant results most likely to reflect valid relationships. While Soar and Soar failed to explore this possibility, they should be commended both for having seen the general problem of spurious significance and for their attempt to cope with this problem.

The Borich, Godbout, Peck, Nash and Poynor (1974) study. This investigation concerned the differential effects of two teacher training programs. One program--the conventional program--employed a traditional cognitive approach to training with emphasis on subject-matter competence and knowledge of teaching methods. The second program--the personalized program--supplemented the traditional cognitive approach with activities intended to encourage the affective development of the trainee. Both conventional and personalized programs included content-related course work, classroom observation and practice teaching. The unique components of the personalized program consisted of (a) repeated counseling sessions with the prospective teacher for engendering attitudinal and personality characteristics thought to be related to effective teaching, (b) the differential assignment of instructional tasks and activities based upon the specific attitudinal and personality characteristics of the prospective teacher, (c) self-observation of affective teaching behavior through videotaping, and (d) affective feedback and assignment of tasks and activities related to this self-observation.

The major research question in the Borich et. al. investigation was whether the two programs had differential effects which varied with the type of individual receiving training. In other words, did the traits and attitudes of the trainees interact with training program? The sample for this investigation consisted of 77 prospective elementary school teachers, 38 receiving teacher training in a conventional program and 39 receiving

teacher training in a personalized program. Trainees were assessed with regard to 22 personality trait and attitude measures upon entry into a training program. Twenty-two trait-program interactions (one for each of the personality trait and attitude measures) were examined with regard to over 100 criteria relevant to the performance of the trainees during practice teaching. Included in the criterion measures were observation measures of the behavior of the teacher trainees during practice teaching as well as evaluations of the trainees' teaching performance by self, by pupils, and by supervisors. Multiple regression analysis techniques were used to perform 2127 significance tests for trait-program interactions with $\alpha = .10$. The regression analysis methods employed are commonly known as trait-treatment interaction methods or aptitude-treatment interaction methods. For a discussion of aptitude-treatment interaction methodology, see Borich, Godbout and Wunderlich (1976).

The entire set of 2127 trait-program interaction results were analyzed with a test equivalent to what has been labeled the NCE test in the present report. A total of 269 of the results were significant at the .10 level, this total exceeding the nonchance expectation of 231. Thus, the entire set of results was not identified as chance-produced. However, for 2127 results with $\alpha = .10$, the average number of significant results expected on the basis of chance (i.e., the chance expectation) was 213. While the 269 obtained significant results seem to reflect some valid relationships, the large majority of these significant results are undoubtedly the work of chance. Given such a situation, it is certainly desirable to attempt to eliminate the majority of the significant results as chance-produced.

Subsets of the 2127 results were analyzed in an attempt to isolate those significant results most likely to be reflective of valid relationships. Three steps were involved in the analysis of subsets of results. The first

step was to form a subset of the results for each criterion instrument (there being 13 criterion instruments involving from 3 to 20 criterion variables). These 13 subsets were then analyzed with a test equivalent to what has been labeled the NCE test in the present report. If there was overall significance for a criterion instrument subset, then the significant results in that subset were retained as being possibly valid. If there was overall nonsignificance for a criterion instrument subset, then any significant results included in that subset were discarded as chance-produced. After the analyses of criterion instrument subsets, 134 of the original 269 significant results were retained as possibly valid while the other 135 significant results were discarded as chance-produced.

In a second step in the analysis of subsets of results, significant criterion instrument subsets were further divided into smaller subsets involving individual criterion variables. Criterion variable subsets each involved 22 results (corresponding to the 22 trait-program interactions) and were analyzed with a test equivalent to what has been labeled the CE test in the present report. If there was overall significance for a criterion variable subset, then the significant results in that subset were retained as possibly valid. If there was overall nonsignificance for a criterion variable subset, then any significant results included in that subset were rejected as chance-produced. After analysis of criterion variable subsets, the 119 significant results associated with 21 of the criterion variables were retained as possibly valid.

The third step involved analysis of subsets of results associated with each of the trait variables. Each trait variable subset involved 21 results, corresponding to the 21 criterion variables which had been found to have significant subsets in the preceding step. Each trait

variable was analyzed with a test equivalent to the CE test. If there was overall significance for a trait variable subset, then the significant results in that subset were retained as possibly valid. If there was overall nonsignificance for a trait variable subset, then the significant results in that subset were discarded as chance-produced. After this last step in the analysis of subsets of results, a total of 107 of the original 269 significant results were retained as valid. These 107 significant results were reported as the findings of the Borich et. al. investigation.

The scheme for analyzing sets of results used by Borich and his coworkers differs from that recommended in the present report. With $\alpha = .10$, these investigators employed the CE test for sets of 21 or more but less than 50 results and the NCE test for sets of 50 or more results. In the present report, the recommendations for $\alpha = .10$ are the CE test for 25 or more but less than 100 results and the NCE test for 100 or more results. Thus, Borich et. al. applied the CE and NCE tests to smaller sets of results than recommended in the present report. The smaller the subsets analyzed, the greater the number of significant results which will be involved in nonsignificant subsets and thereby identified as chance-produced. Thus, Borich et. al. probably rejected a somewhat greater number of significant results than if they had followed the present use recommendations.

A quite consistent pattern is evidenced for the 107 significant trait-program interaction results accepted as valid in the Borich et. al. study. Of these 107 significant results, 96 involved trait and criterion variables both classifiable in terms of social desirability or positiveness. When the trait and criterion variables for these 96 interaction results were both scaled so that high scores were socially desirable or positive, 84 of these results involved a greater (more positive) criterion on trait regression slope for the conventional program than for the personalized

program. Furthermore, 74 of the 96 demonstrated a positive slope for the conventional program but a negative slope for the personalized program.

The 107 significant trait-program interaction results were further examined with regions of significance analyses (Walker & Lev, 1953; Borich, Godbout and Wunderlich, 1976). These analyses demonstrated a tendency for the personalized program to be superior to the conventional program for trainees with less positive traits and attitudes but inferior to the conventional program for trainees with more positive traits and attitudes. These findings suggest that the personal assessment, feedback, and counseling included in the personalized program may be positive experiences for trainees who are less well-adjusted (trainees with less positive traits and attitudes). However, these experiences may be detrimental to self-confident or psychologically well-defensed trainees (trainees with more positive traits and attitudes). These latter trainees may react to the counseling employed in the personalized approach either as a waste of time or as a threat.

This ends our consideration of the Borich et. al. investigation. In summary, this investigation has demonstrated the use of analysis of sets of results in isolating those significant results most likely to reflect valid findings. Such analysis led to the rejection of a majority of the significant results as chance-produced. The significant results retained as valid proved to be highly consistent and suggested a clear difference in the effects of the two training programs.

The Calkins, Godbout, Poynor and Kugle (1976) study. These authors studied fifth grade mathematics achievement as a function of observed teacher behavior. The sample included 32 fifth grade classrooms. Seven measures of mathematics achievement were provided by the 6 subscales and the total score on a mathematics test. This mathematics test was

administered to the pupils in the classrooms at the beginning of a semester (pretest) and then again at the end of that semester (posttest). Seventy-three measures of observed teacher behavior were derived from two classroom observation instruments. The mean and the variablility across observation sessions on each of the 73 teacher behavior measures were calculated separately for each teacher. Four multiple regression analyses were performed for each of the 511 combinations of one of the criterion measures (the 7 mathematics scores) and one of the predictor measures (the 73 teacher behavior measures). These four multiple regression analyses assessed the extent to which gain in pupil mathematics performance was related (1) in linear fashion to mean teacher behavior, (2) in nonlinear fashion to mean teacher behavior, (3) in linear fashion to the variability of teacher behavior, and (4) in nonlinear fashion to the variability of teacher behavior. Each analysis was performed with classroom as the unit of analysis and with $\alpha = .05$.

A total of 2044 analyses (four tests for each of the 511 predictor-criterion combinations) were performed with 71 of these analyses yielding significant results. The authors failed to perform an analysis of the entire set of results and we will now perform such a test. With 2044 results and $\alpha = .05$, the NCE test is recommended and the nonchance expectation is 119 (see Appendix A). The 71 obtained significant results falls far short of the nonchance expectation and it is quite reasonable to assume that all significant results are merely the product of chance. Calkins et. al. came to the same conclusion on the basis of analyses of subsets of results. In summary, this application of methods for analyzing sets of results proved to be extremely valuable, these mehtods exposing that the overall prediction of criterion performance in the Calkins et. al.investigation was not better than chance.

The Good and Grouws (1975) study. Good and Grouws investigated process-

product relationships in 33 fourth grade mathematics classrooms. Product measures included (1) a measure of the gain in mathematics achievement over a school year and (2) a measure of classroom climate (pupil attitude toward teacher and class). The more than 100 process measures included time measures (e.g., time spent in review), observational measures of teacher-student interaction patterns, and measures describing materials and homework assignments. All process-product analyses involved the classroom as the unit of analysis and $\alpha = .10$.

The process-product relationships involving classroom climate were analyzed with product-moment correlation coefficients. The process-product relationships involving mathematics achievement were analyzed in three separate ways. First, product-moment correlation coefficients were calculated based on the entire sample of 33 classrooms. Second, the 33 teachers were rank-ordered according to the mean gain in mathematics achievement demonstrated by their classes. The top 9 teachers were assigned to an extreme group of high effectiveness teachers; and the bottom 9, to an extreme group of low effectiveness teachers. One-way analyses of variance were then employed to assess the mean difference between these high and low effectiveness groups on each of the process measures. Third, similar extreme groups of high and low effectiveness teachers were again formed but with the top and bottom 3 teachers being included. One-way analyses of variance were then performed to test the mean difference between these high and low effectiveness groups on each of the process measures.

Good and Grouws did not perform analyses of sets of results and we will now perform such analyses on their reported results. A total of 429 process-product analyses were reported (including both correlations and analyses of variance). The NCE test is recommended for analysis of a set of 429 results with $\alpha = .10$ and the nonchance expectation is given in Appendix A as 51. The

authors report 81 significant results. Since the number of obtained significant results far exceeds the nonchance expectation, the entire set of results cannot be rejected as chance-produced.

Let us next consider the subset of results for classroom climate. A total of 107 correlations were computed for classroom climate. The NCE test is recommended for analysis of a set of 107 results with $\alpha = .10$ and the nonchance expectation is given in Appendix A as 15. Only 11 results are reported as significant, and we therefore fail to reject the null hypothesis that the entire subset of classroom climate results is chance-produced. Without the benefit of methods for analyzing sets of results, Good and Grouws simply accept the 11 significant results for classroom climate as valid and proceed to devote several pages (Good & Grouws, 1975, pp. 24-27 and pp. 35-36) to description and discussion of these "significant" results.

Next consider the subset of results for mathematics. A total of 322 analyses (including both correlations and analyses of variance) were performed for mathematics achievement. The NCE test is recommended for analysis of this subset of results and the nonchance expectation is given as 40 in Appendix A. A total of 70 significant results are reported for mathematics. The obtained number of significant results far exceeds the nonchance expectation and valid relationships are indicated for mathematics achievement.

Given the overall significance for the set of mathematics achievement results, it is of interest to determine if such significance is attributable to the correlational results for mathematics achievement, the analyses of variance results for extreme mathematics achievement groups, or both. First, consider the subset of correlational results for mathematics achievement. A total of 108 correlations were computed with 12 of these being significant. The NCE test is recommended for 108 results with $\alpha = .10$ and the nonchance expectation is given in Appendix A as 15. The number of obtained significant

results in less than the nonchance expectation, and we therefore accept the null hypothesis that the entire set of correlational results is the product of chance. A different picture is presented by the analysis of variance results for mathematics achievement. 107 analyses of variance were performed both for the 9 member extreme groups and for the 3 member extreme groups yielding a total of 214 results. The NCE test is recommended for analyzing this set of 214 results and the nonchance expectation is given in Appendix A as 28. A total of 58 of these results were found to be significant by Good and Grouws and valid relationships are indicated here.

Without the benefit of analysis of sets of results, Good and Grouws take their individual significant results to indicate valid process-product relationships both for classroom climate and for mathematics achievement. Analyses of sets of results lead to a somewhat different conclusion. Such analyses indicate valid relationships for mathematics achievement but not for classroom climate. Furthermore, analyses of subsets of the mathematics results leads to a more equivocal conclusion concerning valid mathematics achievement relationships. Valid relationships were suggested by the analyses of variance results for the extremes of the sample but not by the correlational results for the entire sample. Considering only the sample extremes, teaching effectiveness seems to have been predictable from process measures. However, considering the entire sample, process measures do not appear to have led to greater than chance prediction of teaching effectiveness.

The Brophy and Evertson (1974a; 1974b) study. These authors conducted a two year, replicated, naturalistic-correlational study of the relationships between process variables and student learning criteria at the second and third grade levels. This study is impressive both because of its size in terms of number of variables studied and because of the care taken in sample selection and data collection.

Only experienced teachers, who had demonstrated relative consistency in the kinds of student learning they had produced over a number of years, were included in the sample. Nineteen teachers served in both years of the study. The total sample for the first year included 31 teachers and the total sample for the second year included 28 teachers. Approximately one-half of the teachers taught in low SES schools and one-half in high SES schools.

Hundreds of process variables (e.g., teacher behavior variables and teacher-student interaction variables) were measured during 4 classroom observations in the first year and 14 classroom observations in the second year. Special care was taken to achieve high reliability and validity for these observational variables which included both low and high inference measures. Low inference process variables were measured separately for morning activities involving the whole class, for afternoon activities involving the whole class, and for activities within reading groups. The student learning criteria (product variables) were residual gain scores for each of the five subtests of the Metropolitan Achievement Test. Two of these subtests involved mathematics achievement while the other three involved language and reading proficiency.

The first step in data analysis involved inspection of the variabilities for the process variables. The variability of each process variable was considered for the low SES sample, the high SES sample, and the total sample. If a process variable had restricted variability for a sample, then that variable was not analyzed with regard to the sample in question. Process variables with adequate variability were analyzed to assess their relationships with product measures. All process-product analyses were performed with the classroom as the unit of analysis and with $\alpha = .10$. Analyses were performed both for linear process-product relationsips and for process-product

relationships involving nonlinearity and interactions with SES. Separate reports present the linear results (Brophy & Evertson, 1974a) and the nonlinear and interaction results (Brophy & Evertson, 1974b). In the present treatment, we will also consider the linear results separately from the nonlinear and interaction results.

The linear results involved product-moment correlation coefficients calculated for each process-product pair. For process-product pairs involving a high inference process variable, six separate correlations were calculated for (1) the high SES sample in Year 1, (2) the low SES sample in Year 1, (3) the total sample in Year 1, (4) the high SES sample in Year 2, (5) the low SES sample in Year 2, and (6) the total sample in Year 2. For process-product pairs involving a low inference process variable, these six correlations were computed separately for (1) morning activities involving the whole class, (2) afternoon activities involving the whole class, and (3) reading group activities. As a result, 18 separate correlations were calculated for each process-product pair involving a low inference process variable. Note that any correlations involving process variables with too restricted a variability were not reported. The total number of process-product correlations actually reported by Brophy and Evertson was 12977.

Brophy and Evertson did not perform analyses of sets of results and we will now perform such analyses for their linear results. The NCE test is recommended for analyzing the entire set of 12977 correlations. However, this test cannot be applied because we have not calculated nonchance expectations for sets involving more than 10000 results (see Appendix A). The CE test can be applied to this set of results since the chance expectation is simply the number of results times α. Application of the CE test rather than the NCE test will lead to both higher statistical power and a higher

Type I error rate. Brophy and Evertson report 1219 of the 12977 correlations to be significant at the .10 level. The chance expectation is 1297.7 and the obtained number of significant results does not exceed this chance expectation. The CE test therefore leads to the conclusion that the entire set of linear resuts is the product of chance.

A comment is in order concerning the power of the CE test when applied to such a large set of results. When applied to 5000 minimally dependent results, the CE test power (1) exceeds 0.80 when 0.4% or more of the results reflect strong population relationships[20] (from Table 10), (2) exceeds 0.90 when 0.6% or more of the results reflect strong population relationships (from Table 12), and (3) exceeds 0.95 when 0.8% or more of the results reflect strong population relationships (from Table 13). The power of the CE test as applied to a set of 12977 results which are not minimally dependent will be even higher, since power increases both as the number of results in a set increases and as the dependence among the results increases. Thus, the statistical power achieved in analyzing the set of 12977 results with the CE test is extremely high. Surely the existence of about 0.5% strong population relationships or 1% moderate population relationships[21] should have been detected with almost complete certainty.

Given nonsignificance for the entire set of linear results, consideration of these results would usually end at this point. However, we have decided to make an exception and present analyses of subsets of the Brophy and Evertson linear results. The Brophy and Evertson study involved several

---

[20] A strong population relationship is one which will always lead to a significant sample result.

[21] A moderate population relationship is one which will lead to a significant sample results approximately 50% of the time.

different types of process variables. There is some possibility that several valid relationships existed for one type of process measure, that these valid relationships represented a very small percentage (e.g., less than 0.5%) of the total set of linear results, and that these valid relationships therefore went undetected in the analysis of the total set. To check this possibility, we performed analyses of subsets corresponding to different types of process measures.

Brophy and Evertson (1974a) present their linear results in six separate tables, each table corresponding to a different type of process measure. For example, the first table involved high inference process measures based upon a set of classroom observation scales suggested by Emmer and Peck (1973), while the sixth table involved low inference process measures based upon the Expanded Brophy-Good Dyadic Observation System (Brophy and Good, 1970). For the present report, the subset of Year 1 results and the subset of Year 2 results were analyzed separately for each of the six tables presented by Brophy and Evertson. The analyses of these subsets are presented in Table 30. Note that all subset analyses involved the NCE test and $\alpha = .10$.

As shown in Table 30, two of the 12 subsets of results did display overall significance--the Table II results for Year 2 and the Table VI results for Year 1. The importance of these significant subsets is lessened by the fact that neither the Table II results nor the Table VI results achieved overall significance for both years. In summary then, analyses of subsets of the linear results in the Brophy and Evertson study provide some evidence for valid relationships involving two different types of process measures. However, in both cases, such evidence was limited to one of the two years of the investigation. This conclusion contrasts sharply with Brophy and Evertson's conclusions concerning their linear results. These investigators did not analyze sets of results and rather described and

Table 30.

NCE Test Analyses of Subsets of the Linear Results Reported by Brophy and Evertson (1974a)

	YEAR 1				YEAR 2			
Table in Brophy & Evertson Report	# results in subset	# obtained significant results	nonchance expectation	is subset significant ?	# results in subset	# obtained significant results	nonchance expectation	is subset significant ?
I*	180	8	24	no	180	15	24	no
II	869	38	98	no	900	103	103	yes
III	629	61	73	no	630	50	74	no
IV	475	54	56	no	500	38	60	no
V	130	17	18	no	405	38	49	no
VI	3521	403	374	yes	4558	394	462	no

Note: $\alpha = .10$ in all cases.

*The table numbers in Brophy and Evertson (1974a) are Arabic numerals (e.g., Table 1). In referring to these tables, we have substituted the corresponding Roman numerals to avoid confusion with our own table numbers.

interpreted each individual "significant" correlation as if it reflected a valid relationship.

We will now consider the nonlinear and interaction results reported by Brophy and Evertson (1974b). The nonlinear and interaction analyses performed by these authors involved the same process and product variables previously considered with regard to linear relationships. The nonlinear and interaction analyses included (1) a test for curvilinear (quadratic) interaction between the process variable and SES (high vs. low) with regard to the product variable, (2) a test for common curvilinear relationship between the process variable and the product variable across the two SES samples, and (3) a test for linear interaction between the process variable and SES with regard to the product variable. All of these analysis involved the classroom as the unit of analyses and $\alpha = .10$. Note that the results of any analysis involving a process variable with too restricted a variability were not reported.

With regard to their nonlinear and interaction results, Brophy and Evertson did not analyze sets of results, and we will now perform such analyses. A problem arises because Brophy and Evertson only reported selected nonlinear and interaction results rather than the total set of such results. For this reason, we cannot consider the total set of analyses performed by these investigators. Instead we must limit our attention to those analyses reported as significant or nonsignificant in Brophy and Evertson (1974b). Counting the number of significant and nonsignificant results in the Brophy and Evertson report proved to be quite difficult and the methods we used to make such counts are presented in Appendix E.

The total number of nonlinear and interaction analyses with a known outcome (i.e., significance or nonsignificance) is 12739 and of these 1200 were significant ($\alpha = .10$). The NCE test is recommended for analyzing

the entire set of these 12739 results. However, this test cannot be applied because we have not calculated nonchance expectations for sets involving more than 10000 results (see Appendix A). Therefore, the CE test will be applied instead of the NCE test. The chance expectation for the set of 12739 results is 1273.9 and the 1200 obtained significant results do not exceed this expectation. The CE test thus leads to the conclusion that this entire set of nonlinear and interaction results is the product of chance.

Given the overall nonsignificance for the entire set of nonlinear and interaction results and given the high statistical power achieved in a test of such a large number of dependent results, consideration of subsets of these results would not usually be entertained. However, as was done for the linear results, we will make an exception and analyze subsets of the nonlinear and interaction results, these subsets being defined on the basis of different types of process measures.

Brophy and Evertson (1974b) present their nonlinear and interaction results in 10 separate tables, each table corresponding to a different type of process measure. For the present report, we have separately analyzed the subsets of results corresponding to each of these 10 tables. The outcomes of analyses of these subsets are presented in Table 31. Note that all of these analyses involved the NCE test and $\alpha = .10$.

As shown in Table 31, eight of the subsets were not significant, leading to the conclusion that these subsets of results were the products of chance. Two of the ten subsets did display overall significance--the Table V results and the Table VIII results. The resulting conclusion is that, of the total of 1200 significant nonlinear and interaction results, only the 265 significant results for Tables V and VIII are likely to reflect any concentration of valid relationships. This conclusion provides a sharp contrast to Brophy and Everton's conclusion for their nonlinear and

Table 31.

NCE Test Analyses of Subsets of the Nonlinear and Interaction Results Reported by Brophy and Evertson (1974b)

Table in Brophy & Evertson Report	# results in subset	# obtained significant results	nonchance expectation	is subset significant ?
I*	657	54	76	no
II	1348	123	148	no
III	1120	111	124	no
IV	324	32	40	no
V	301	41	38	yes
VI	2130	227	231	no
VII	1930	154	210	no
VIII	1989	224	216	yes
IX	2462	197	265	no
X	478	37	56	no

Note: $\alpha = .10$ in all cases.

*The table numbers in Brophy and Evertson (1974b) are Arabic numerals (e.g., Table 1). In referring to these tables, we have substituted the corresponding Roman numerals to avoid confusion with our own table numbers.

interaction results. These investigators described and interpreted all 1200 individual "significant" results as if each reflected a valid relationship.

In summary then, there is no strong evidence for better than chance prediction of criterion performance in either Brophy and Evertson's linear results or in their nonlinear and interaction results. Analyses of the entire set of linear results and of the entire set of nonlinear and interaction results both failed to achieve significance. Only when smaller subsets of the results were considered did some evidence of valid relationships appear.

Brophy and Evertson would disagree with these conclusions derived from analyses of sets of their results, these authors having suggested that such analyses are inappropriate. Consider the following quote (Brophy & Evertson, 1974b, p. 169):

> The upshot of all this is that the real probability of a given finding is unknown and for all intents and purposes unknowable. This also goes for the probability of finding a given percentage of our relationships to be "statistically significant," regardless of what significance level is chosen as "significant." Given that a very large number of variables were included in an attempt to study "anything" which might be related to student learning, that many variables had no data, no variance, and/or drastically abnormal distributions, and that many variables had ipsative relationships with one another, it simply is not possible to ascertain the probability level of a given finding or to estimate with any confidence the percentage of findings which "should be expected" to be statistically significant at a given probability level.

These comments center around three problems--(1) the fact that a large number of variables were included, (2) the nonnormal distributions for many of the variables, and (3) the dependence among the results. From our perspective, the inclusion of a large number of variables would not preclude analysis of sets of results. In fact, we have argued repeatedly in the present report that analysis of sets of results is indispensible if large-scale multivariate exploratory research is to be meaningfully evaluated. Also, our simulation results have indicated that analysis of sets of results can be meaningfully performed regardless of the dependence among the results. This leaves only

the problem of nonnormality to consider. Nonnormality presents a more difficult problem which cannot be so easily dispensed with.

If the assumptions underlying a statistical test are violated, then the results produced by use of that test need not have α as the chance probablility for significance--i.e., the Type I error rate for use of that test may not equal α. If the chance probability of significance for an individual result is not known, then the CE and NCE tests are no longer applicable. Only when the chance probability for an individual result is known can we determine the values for the chance expectation and the nonchance expectation.

An underlying assumption for testing the significance of a product-moment correlation coefficient is that the two variables be distributed in bivariate normal fashion. If the variables are not so distributed, then obtained correlation coefficients will be restricted in magnitude (Hays, 1963; McNemar, 1969; Calkins, 1974). Given a population correlation of 0 and nonnormally distributed variables, the restricted magnitudes of obtained sample correlations may yield a probability less than α of chance significance. For such results, chance expectations and nonchance expectations, based upon the assumption that the chance probability of significance equals α, may be too high, and the CE and NCE tests may reject sets of results which actually reflect a high concentration of valid relationships.

If the bivariate distributions are not normal, then the ability of the CE and NCE tests to detect sets of bivariate correlation results involving valid relationships may be reduced. Of crucial importance is the extent of this reduction in the detection power of the CE and NCE tests. If this reduction is small (i.e., if the significance test for a correlation is robust under conditions of nonnormality), then application of the CE and NCE tests will still lead to reasonably valid conclusions. However, if this

reduction is large (i.e., if the significance test for a correlation is not robust under conditions of nonnormality), then application of the CE and NCE tests will lead to completely invalid conclusions.

We have not found information concerning the robustness of the significance test for a correlation coefficient when the assumption of normality has been violated. However, much information is available for the t-test and the analysis of variance. For these latter tests, violation of the assumption of normality is almost of no importance (McNemar, 1962; Glass & Stanley, 1970). For example, Glass and Stanley conclude that, even with nonnormal data, the Type I error rate for these tests remains almost exactly at the value specified by the researcher, namely α. Thus, nonnormal data should prove little problem for the application of the CE and NCE tests to sets of t-test or analysis of variance results.

It may not be unreasonable to assume that the robustness of the significance test for a correlation coefficient is somewhat similar to that for the t-test or the analysis of variance. All of these significance tests are examples of the general multiple regression analysis (Cohen, 1969) and are likely to have similar properties. A reasonably safe approach to analyzing sets of correlation results involving cases of nonnormality would be to illiminate those variables which represent extreme departures from normality and then apply the CE and NCE tests to the remaining results.

Considering analysis of sets of Brophy and Evertson's results, one might attribute the failure to detect a large number of valid relationships to the effects of nonnormality. However, recall that these investigators did not report results for measures with the most restricted variabilities. They have illiminated the most glaring examples of nonnormality and this action should have lessened the problem of nonnormality in their reported results. Furthermore, there is additional evidence that the Brophy and Evertson results

do not reflect a large number of valid relationships. Recall that the Brophy and evertson study involved replication over two years. Had a large number of valid relationships been present, then we would expect to find a good number of correlations to be significant in the same direction for both years. Let us consider this matter with regard to the Brophy and Evertson linear results.

Reconsider the Year 1 results in Table 30 for a moment. The highest proportion of significant results in Year 1 occurred for the Table VI results. Of all of the Year 1 significant results, those for Table VI should have had the greatest tendency to be significant again in Year 2. We therefore examined the Table VI results across the two years to determine the number of these correlations which were significant in the same direction for both years.

Only 3169 of the Table VI correlations were actually reported for both years. Of these 3169 correlations, we counted only 24 which were significant in the same direction for both years. Obviously, a large number of valid relationships are not indicated here.

Although a large number of valid relationships do not appear to be involved in the Year 1 results for Table VI, the obtained number of significant results exceeded chance expectation. This suggests that nonnormality had little effect with regard to the Table VI results for Year 1. Had nonnormality been a factor, then the obtained number of significant results should have been far below chance expectation, especially given that few valid relationships appear to be present. This limited evidence suggests that the effect of violations of the assumption of normality in Brophy and Evertson's reported results probably had only a minor effect.

Conclusion. These example applications of the CE and NCE tests to classroom education research indicate that a majority of the "significant"

findings in that research may be spurious. In the five investigations considered in this report, valid relationships appear to be few and scattered. The present methods for analyzing sets of results lead us to question the value of the exploratory, shotgun approach to studying classroom education.

Certainly, valid relationships must exist between pupil outcome and classroom events. The issue is not whether such relationships exist but rather why these relationships have not been more readily detected by research. Perhaps, the relatively simple analyses performed in most research are not sensitive to the majority of the valid relationships which may be of a highly complex multivariate nature. Also, the typical observational measures of classroom behavior may be too errorful to allow relationships to be detected. An appropriate direction for future classroom education research may be (1) the development of more reliable and valid measures of classroom activity and (2) the use of such measures in an attempt to identify complex multivariate relationships with pupil outcome.

References

Bock, D. R. Multivariate statistical methods in behavioral research. New York: McGraw-Hill, 1975.

Borich, G. D., & Godbout, R. C. Extreme groups designs and the calculation of statistical power. Educational and Psychological Measurement, 1974, 34, 663-675.

Borich, G. D., Godbout, R. C., Peck, R. F., Kash, M. M., & Poynor, L. H. The evaluation of personalized teacher training. Final report to the National Institute of Education, C-74-0088. Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1974.

Borich, G. D., Godbout, R. C., & Wunderlich, K. W. The analysis of aptitude-treatment interactions: Computer programs and calculations. Chicago: International Educational Services, 1976.

Bradley, J. V. Distribution-free statistical tests. Englewood Cliffs, NJ: Prentice-Hall, 1968.

Bradley, J. V. A survey of sign tests based on the binomial distribution. Journal of Quality Technology, 1969, 1, 89-101.

Brophy, J., & Good, T. The Brophy-Good dyadic interaction system. In A. Simon and E. Boyer (Eds.), Mirrors for behavior: An anthology of observation instruments continued, 1970 supplement (Vol. A). Philadelphia: Research for Better Schools, Inc., 1970.

Brophy, J., & Evertson, C. Process-product correlations in the Texas Teacher Effectiveness Project: Final report. (Research Report No. 74-4). Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1974a.

Brophy, J., & Evertson, C. The Texas Teacher Effectiveness Project: Presentation of non-linear relationships and summary discussion. (Research Report No. 74-6). Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1974b.

Calkins, D. S. Some effects of non-normal distribution on the magnitude of the Pearson product moment correlation coefficient. Interamerican Journal of Psychology, 1974, 8, 261-288.

Calkins, D. S., Godbout, R. C., Poynor, L. H., & Kugle, C. L. Relationships between pupil achievement and characteristics of observed teacher behavior distributions. (Research Report No. 6). Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1976.

Cohen, J. Multiple regression as a general data analytic system. Psychological Bulletin, 1968, 70, 426-443.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Cooley, W. W., & Lohnes, P. R. Multivariate data analysis. New York: John Wiley & Sons, 1971.

Edwards, A. L. Experimental design in psychological research. New York: Holt, Rinehart and Winston, 1968.

Emmer, E., & Peck, R. Dimensions of classroom behavior. Journal of Educational Psychology, 1973, 64, 223-240.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Godbout, R. C. The problem of chance significant findings in educational research. Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1975.

Good, T. L., & Grouws, D. A. Process-product relationship in fourth grade mathematics classrooms. Final report to the National Institute of Education, Washington, D.C. Columbia: The University of Missouri, 1975.

Guertin, W. H., & Bailey, J. P. Introduction to modern factor analysis. Ann Arbor, Michigan: Edwards Brothers, 1970.

Hays, W. L. Statistics. New York: Holt, Rinehart and Winston, 1963.

Kash, M. M., Borich, G. D., & Fenton, K. S. Final report: Teacher behavior and pupil self-concept. Austin, Texas: The Research and Development Center for Teacher Education, The University of Texas, 1976.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart and Winston, 1973.

Miller, R. G. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

McNemar, Q. Psychological statistics (3rd ed.). New York: Wiley, 1962.

McNemar, Q. Psychological statistics (4th ed.). New York: Wiley, 1969.

Poynor, H. Selecting units of analyses. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, NJ: Educational Technology Publications, 1974.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley and Sons, 1973.

Rosenshine, B. Teaching behaviors and student achievement. London: International Association for the Evaluation of Educational Achievement, 1971.

Simon, A., & Boyer, E. G. (Eds.). Mirrors for behavior: An anthology of observation instruments. Philadelphia: Research for Better Schools, Inc., 1970.

Soar, R. S., & Soar, R. M. Classroom behavior, pupil characteristics and pupil growth for the school year and the summer. Institute for Development of Human Resources. Gainesville: The University of Florida, 1973.

Walker, H., & Lev, J. Statistical inference. New York: Holt, Rinehart & Winston, 1953.

Reference Notes

1. International Mathematical and Statistical Libraries, Suite 510, 6200 Hillcroft, Houston, Texas, 77036.

Nonchance Expectation Values for Varying Sized Sets of Results (N = Number of Results in a Set of Results) and for Varying α.

N	α .01	.05	.10	.25	N	α .01	.05	.10	.25
1	1	1	1	1	51	4	6	9	16
2	2	2	2	2	52	4	6	9	16
3	2	2	2	2	53	4	6	9	16
4	2	2	2	3	54	4	7	9	17
5	2	2	2	3	55	4	7	9	17
6	2	2	3	3	56	4	7	10	17
7	2	2	3	3	57	4	7	10	17
8	2	3	3	4	58	4	7	10	18
9	2	3	3	4	59	4	7	10	18
10	2	3	3	4	60	4	7	10	18
11	2	3	3	5	61	4	7	10	18
12	2	3	4	5	62	4	7	10	19
13	2	3	4	5	63	4	7	10	19
14	2	3	4	6	64	4	7	11	19
15	2	3	4	6	65	4	7	11	20
16	3	3	4	6	66	4	7	11	20
17	3	4	4	6	67	4	7	11	20
18	3	4	4	7	68	4	8	11	20
19	3	4	5	7	69	4	8	11	21
20	3	4	5	7	70	4	8	11	21
21	3	4	5	8	71	4	8	11	21
22	3	4	5	8	72	4	8	12	21
23	3	4	5	8	73	4	8	12	22
24	3	4	5	8	74	4	8	12	22
25	3	4	5	9	75	4	8	12	22
26	3	4	6	9	76	4	8	12	23
27	3	4	6	9	77	4	8	12	23
28	3	4	6	9	78	4	8	12	23
29	3	5	6	10	79	4	8	12	23
30	3	5	6	10	80	4	8	13	24
31	3	5	6	10	81	4	8	13	24
32	3	5	6	11	82	4	9	13	24
33	3	5	7	11	83	4	9	13	24
34	3	5	7	11	84	5	9	13	25
35	3	5	7	11	85	5	9	13	25
36	3	5	7	12	86	5	9	13	25
37	3	5	7	12	87	5	9	13	25
38	3	5	7	12	88	5	9	13	26
39	3	5	7	13	89	5	9	14	26
40	3	5	7	13	90	5	9	14	26
41	3	6	8	13	91	5	9	14	26
42	3	6	8	13	92	5	9	14	27
43	3	6	8	14	93	5	9	14	27
44	3	6	8	14	94	5	9	14	27
45	4	6	8	14	95	5	9	14	28
46	4	6	8	14	96	5	10	14	28
47	4	6	8	15	97	5	10	15	28
48	4	6	9	15	98	5	10	15	28
49	4	6	9	15	99	5	10	15	29
50	4	6	9	16	100	5	10	15	29

N	α .01	.05	.10	.25
110	5	10	16	32
120	5	11	17	34
130	6	12	18	37
140	6	12	20	39
150	6	13	21	42
160	6	14	22	45
170	6	14	23	47
180	6	15	24	50
190	7	16	25	52
200	7	16	27	55
210	7	17	28	58
220	7	18	29	60
230	7	18	30	63
240	8	19	31	65
250	8	19	32	68
260	8	20	33	71
270	8	21	34	73
280	8	21	36	76
290	8	22	37	78
300	9	22	38	81
310	9	23	39	84
320	9	24	40	86
330	9	24	41	89
340	9	25	42	91
350	9	25	43	94
360	10	26	44	96
370	10	27	45	99
380	10	27	47	102
390	10	28	48	104
400	10	28	49	107
410	10	29	50	109
420	11	30	51	112
430	11	30	52	115
440	11	31	53	117
450	11	31	54	120
460	11	32	55	122
470	11	33	56	125
480	12	33	58	127
490	12	34	59	130
500	12	34	60	132
510	12	35	61	135
520	12	35	62	138
530	12	36	63	140
540	12	37	64	143
550	13	37	65	145
560	13	38	66	148
570	13	38	67	150
580	13	39	68	153
590	13	39	69	156
600	13	40	70	158

N	α .01	.05	.10	.25
610	13	41	72	161
620	14	41	73	163
630	14	42	74	166
640	14	42	75	168
650	14	43	76	171
660	14	43	77	173
670	14	44	78	176
680	14	45	79	179
690	15	45	80	181
700	15	46	81	184
710	15	46	82	186
720	15	47	83	189
730	15	47	84	191
740	15	48	86	194
750	15	49	87	196
760	16	49	88	199
770	16	50	89	202
780	16	50	90	204
790	16	51	91	207
800	16	51	92	209
810	16	52	93	212
820	16	53	94	214
830	17	53	95	217
840	17	54	96	219
850	17	54	97	222
860	17	55	98	225
870	17	55	99	227
880	17	56	100	230
890	17	56	102	232
900	18	57	103	235
910	18	58	104	237
920	18	58	105	240
930	18	59	106	242
940	18	59	107	245
950	18	60	108	247
960	18	60	109	250
970	19	61	110	253
980	19	61	111	255
990	19	62	112	258
1000	19	63	113	260
1025	19	64	116	267
1050	20	65	119	273
1075	20	67	121	279
1100	20	68	124	286
1125	21	70	126	292
1150	21	71	129	298
1175	21	72	132	305
1200	22	74	134	311
1225	22	75	137	317
1250	22	76	140	324

N	α .01	.05	.10	.25
1275	23	78	142	330
1300	23	79	145	336
1325	23	81	148	343
1350	24	82	150	349
1375	24	83	153	356
1400	24	85	155	362
1425	25	86	158	368
1450	25	87	161	375
1475	25	89	163	381
1500	26	90	166	387
1525	26	91	169	394
1550	26	93	171	400
1575	27	94	174	406
1600	27	96	176	413
1625	27	97	179	419
1650	28	98	182	425
1675	28	100	184	432
1700	28	101	187	438
1725	29	102	190	444
1750	29	104	192	451
1775	29	105	195	457
1800	29	106	197	463
1825	30	108	200	470
1850	30	109	203	476
1875	30	111	205	482
1900	31	112	208	489
1925	31	113	210	495
1950	31	115	213	501
1975	32	116	216	508
2000	32	117	218	514
2025	32	119	221	520
2050	33	120	223	527
2075	33	121	226	533
2100	33	123	229	539
2125	34	124	231	546
2150	34	125	234	552
2175	34	127	237	558
2200	35	128	239	565
2225	35	129	242	571
2250	35	131	244	577
2275	35	132	247	584
2300	36	133	250	590
2325	36	135	252	596
2350	36	136	255	603
2375	37	137	257	609
2400	37	139	260	615
2425	37	140	263	622
2450	38	141	265	628
2475	38	143	268	634
2500	38	144	270	641

N	α .01	.05	.10	.25
2525	39	146	273	647
2550	39	147	275	653
2575	39	148	278	660
2600	39	150	281	666
2625	40	151	283	672
2650	40	152	286	678
2675	40	154	288	685
2700	41	155	291	691
2725	41	156	294	697
2750	41	158	296	704
2775	42	159	299	710
2800	42	160	301	716
2825	42	162	304	723
2850	43	163	307	729
2875	43	164	309	735
2900	43	166	312	742
2925	43	167	314	748
2950	44	168	317	754
2975	44	170	320	761
3000	44	171	322	767
3025	45	172	325	773
3050	45	174	327	780
3075	45	175	330	786
3100	46	176	332	792
3125	46	178	335	799
3150	46	179	338	805
3175	46	180	340	811
3200	47	182	343	817
3225	47	183	345	824
3250	47	184	348	830
3275	48	186	351	836
3300	48	187	353	843
3325	48	188	356	849
3350	49	189	358	855
3375	49	191	361	862
3400	49	192	364	868
3425	49	193	366	874
3450	50	195	369	881
3475	50	196	371	887
3500	50	197	374	893
3525	51	199	376	900
3550	51	200	379	906
3575	51	201	382	912
3600	52	203	384	918
3625	52	204	387	925
3650	52	205	389	931
3675	52	207	392	937
3700	53	208	394	944
3725	53	209	397	950
3750	53	211	400	956

N	α .01	.05	.10	.25
3775	54	212	402	963
3800	54	213	405	969
3825	54	215	407	975
3850	55	216	410	982
3875	55	217	413	988
3900	55	219	415	994
3925	55	220	418	1001
3950	56	221	420	1007
3975	56	223	423	1013
4000	56	224	425	1019
4025	57	225	428	1026
4050	57	227	431	1032
4075	57	228	433	1038
4100	58	229	436	1045
4125	58	231	438	1051
4150	58	232	441	1057
4175	58	233	443	1064
4200	59	234	446	1070
4225	59	236	449	1076
4250	59	237	451	1082
4275	60	238	454	1089
4300	60	240	456	1095
4325	60	241	459	1101
4350	60	242	461	1108
4375	61	244	464	1114
4400	61	245	467	1120
4425	61	246	469	1127
4450	62	248	472	1133
4475	62	249	474	1139
4500	62	250	477	1146
4525	63	252	479	1152
4550	63	253	482	1158
4575	63	254	485	1164
4600	63	256	487	1171
4625	64	257	490	1177
4650	64	258	492	1183
4675	64	260	495	1190
4700	65	261	497	1196
4725	65	262	500	1202
4750	65	263	503	1209
4775	65	265	505	1215
4800	66	266	508	1221
4825	66	267	510	1227
4850	66	269	513	1234
4875	67	270	515	1240
4900	67	271	518	1246
4925	67	273	521	1253
4950	67	274	523	1259
4975	68	275	526	1265
5000	68	277	528	1272

N	α .01	.05	.10	.25
5025	68	278	531	1278
5050	69	279	533	1284
5075	69	281	536	1291
5100	69	282	539	1297
5125	70	283	541	1303
5150	70	284	544	1309
5175	70	286	546	1316
5200	70	287	549	1322
5225	71	288	551	1328
5250	71	290	554	1335
5275	71	291	557	1341
5300	72	292	559	1347
5325	72	294	562	1354
5350	72	295	564	1360
5375	72	296	567	1366
5400	73	298	569	1372
5425	73	299	572	1379
5450	73	300	574	1385
5475	74	302	577	1391
5500	74	303	580	1398
5525	74	304	582	1404
5550	74	305	585	1410
5575	75	307	587	1417
5600	75	308	590	1423
5625	75	309	592	1429
5650	76	311	595	1435
5675	76	312	598	1442
5700	76	313	600	1448
5725	76	315	603	1454
5750	77	316	605	1461
5775	77	317	608	1467
5800	77	319	610	1473
5825	78	320	613	1479
5850	78	321	615	1486
5875	78	322	618	1492
5900	78	324	621	1498
5925	79	325	623	1505
5950	79	326	626	1511
5975	79	328	628	1517
6000	80	329	631	1524
6025	80	330	633	1530
6050	80	332	636	1536
6075	80	333	639	1542
6100	81	334	641	1549
6125	81	336	644	1555
6150	81	337	646	1561
6175	82	338	649	1568
6200	82	339	651	1574
6225	82	341	654	1580
6250	82	342	656	1587

N	α .01	.05	.10	.25
6275	83	343	659	1593
6300	83	345	662	1599
6325	83	346	664	1605
6350	84	347	667	1612
6375	84	349	669	1618
6400	84	350	672	1624
6425	84	351	674	1631
6450	85	353	677	1637
6475	85	354	680	1643
6500	85	355	682	1650
6525	86	356	685	1656
6550	86	358	687	1662
6575	86	359	690	1668
6600	86	360	692	1675
6625	87	362	695	1681
6650	87	363	697	1687
6675	87	364	700	1694
6700	88	366	703	1700
6725	88	367	705	1706
6750	88	368	708	1712
6775	88	370	710	1719
6800	89	371	713	1725
6825	89	372	715	1731
6850	89	373	718	1738
6875	90	375	720	1744
6900	90	376	723	1750
6925	90	377	726	1757
6950	90	379	728	1763
6975	91	380	731	1769
7000	91	381	733	1775
7025	91	383	736	1782
7050	92	384	738	1788
7075	92	385	741	1794
7100	92	386	743	1801
7125	92	388	746	1807
7150	93	389	749	1813
7175	93	390	751	1819
7200	93	392	754	1826
7225	94	393	756	1832
7250	94	394	759	1838
7275	94	396	761	1845
7300	94	397	764	1851
7325	95	398	766	1857
7350	95	399	769	1863
7375	95	401	772	1870
7400	96	402	774	1876
7425	96	403	777	1882
7450	96	405	779	1889
7475	96	406	782	1895
7500	97	407	784	1901

N	α .01	.05	.10	.25
7525	97	409	787	1908
7550	97	410	789	1914
7575	98	411	792	1920
7600	98	413	795	1926
7625	98	414	797	1933
7650	98	415	800	1939
7675	99	416	802	1945
7700	99	418	805	1952
7725	99	419	807	1958
7750	100	420	810	1964
7775	100	422	812	1970
7800	100	423	815	1977
7825	100	424	818	1983
7850	101	426	820	1989
7875	101	427	823	1996
7900	101	428	825	2002
7925	102	429	828	2008
7950	102	431	830	2014
7975	102	432	833	2021
8000	102	433	835	2027
8025	103	435	838	2033
8050	103	436	841	2040
8075	103	437	843	2046
8100	104	439	846	2052
8125	104	440	848	2059
8150	104	441	851	2065
8175	104	442	853	2071
8200	105	444	856	2077
8225	105	445	858	2084
8250	105	446	861	2090
8275	106	448	864	2096
8300	106	449	866	2103
8325	106	450	869	2109
8350	106	452	871	2115
8375	107	453	874	2121
8400	107	454	876	2128
8425	107	455	879	2134
8450	107	457	881	2140
8475	108	458	884	2147
8500	108	459	887	2153
8525	108	461	889	2159
8550	109	462	892	2165
8575	109	463	894	2172
8600	109	464	897	2178
8625	109	466	899	2184
8650	110	467	902	2191
8675	110	468	904	2197
8700	110	470	907	2203
8725	111	471	909	2209
8750	111	472	912	2216

N	α .01	.05	.10	.25
8775	111	474	915	2222
8800	111	475	917	2228
8825	112	476	920	2235
8850	112	477	922	2241
8875	112	479	925	2247
8900	113	480	927	2254
8925	113	481	930	2260
8950	113	483	932	2266
8975	113	484	935	2272
9000	114	485	938	2279
9025	114	487	940	2285
9050	114	488	943	2291
9075	114	489	945	2298
9100	115	490	948	2304
9125	115	492	950	2310
9150	115	493	953	2316
9175	116	494	955	2323
9200	116	496	958	2329
9225	116	497	961	2335
9250	116	498	963	2342
9275	117	500	966	2348
9300	117	501	968	2354
9325	117	502	971	2360
9350	118	503	973	2367
9375	118	505	976	2373
9400	118	506	978	2379
9425	118	507	981	2386
9450	119	509	983	2392
9475	119	510	986	2398
9500	119	511	989	2404
9525	120	512	991	2411
9550	120	514	994	2417
9575	120	515	996	2423
9600	120	516	999	2430
9625	121	518	1001	2436
9650	121	519	1004	2442
9675	121	520	1006	2448
9700	121	522	1009	2455
9725	122	523	1011	2461
9750	122	524	1014	2467
9775	122	525	1017	2474
9800	123	527	1019	2480
9825	123	528	1022	2486
9850	123	529	1024	2492
9875	123	531	1027	2499
9900	124	532	1029	2505
9925	124	533	1032	2511
9950	124	535	1034	2518
9975	125	536	1037	2524
10000	125	537	1040	2530

Appendix B

Validity of the Simulation Results

Each reported simulation involved a set of population intercorrelation values and corresponding sets of values simulated for a number of samples. A simulation is valid to the extent that simulated sample values differed from the corresponding population values within the range of sampling error.

Notation: In the present appendix, the following notations will be used:

n_P = the number of predictor variables involved in the simulation;

n_C = the number of criterion variables involved in the simulation;

n_r = the number of intercorrelations under consideration;

$n_{r(Tot)} = (n_P + n_C)(n_P + n_C - 1)/2$, the total number of intercorrelations involved in the simulation;

$n_{r(P)} = (n_P)(n_P - 1)/2$, the number of P matrix (predictor-predictor) intercorrelations involved in the simulation;

$n_{r(C)} = (n_C)(n_C - 1)/2$, the number of C matrix (criterion-criterion) intercorrelations involved in the simulation;

$n_{r(P-C)} = (n_P)(n_C)$, the number of P-C matrix (predictor-criterion) intercorrelations involved in the simulation;

n_s = the number of samples in the simulation;

Appendix B (Cont.)

N_S = the number of subjects in a sample;

R_i = the population value for the i^{th} intercorrelation;

r_{ij} = the simulated value for the i^{th} intercorrelation in the j^{th} sample;

$\bar{r}_i = \sum_{j=1}^{n_s} r_{ij}$, the average simulated value across samples for the i^{th} intercorrelation;

$\bar{\bar{r}} = \sum_{j=1}^{n_s} \sum_{i=1}^{n_r} r_{ij}$, the grand mean of the simulated correlation values;

$\sigma_{r_{ij}} = \dfrac{1 - R_i^2}{\sqrt{N_S}}$, the expected (population) value for the sampling error of the i^{th} intercorrelation in the j^{th} sample--adapted from McNemar (1962);

$\sigma_{\bar{r}_i} = \dfrac{\sigma_{r_{ij}}}{\sqrt{n_s}}$, the expected (population) value for the sampling error of the mean value across samples for the i^{th} intercorrelation--based upon familiar relationship, $\sigma_{\bar{x}} = \sigma / \sqrt{N}$;

$s_{\bar{r}} = \dfrac{\sum_{i=1}^{n_r} (\bar{r}_i - \bar{\bar{r}})^2}{n_r - 1}$, the unbiased estimate of the variability of the average simulated intercorrelations;

$e_i = \bar{r}_i - R_i$, the obtained sampling error for the i^{th} mean intercorrelation;

$\bar{e} = \sum_{i=1}^{n_r} e_i$, the average sampling error for the mean intercorrelations;

Appendix B (Cont.)

and $s_e = \frac{\sum_{i=1}^{n_r} (e_i - \bar{e})^2}{n_r - 1}$, the unbiased estimate of the variability of the sampling errors for the mean intercorrelations.

The simulation program did not report the individual simulated values for each sample--i.e., the r_{ij} values. Rather the program reported only the mean simulated values across samples--i.e., the $\bar{r}_i$ values. Simulation validity was assessed by comparing statistics ($\bar{\bar{r}}$, $s_{\bar{r}}$, $\bar{e}$ and s_e) based upon the $\bar{r}_i$ values to corresponding population values.

Types of simulations reported. All of the reported simulations involved Type I error rates. For each simulation, the population values for all predictor-criterion (P-C matrix) intercorrelations were always 0. However, the simulations did differ wtih regard to the population values for the predictor-predictor (P matrix) and criterion-criterion (C matrix) intercorrelations and therefore with regard to the dependence among the simulated predictor-criterion results. One series of simulations involved minimal dependence--i.e., all predictor-predictor and criterion-criterion population intercorrelations were 0. A second series of simulations involved uniform correlated dependence--i.e., all predictor-predictor and criterion-criterion intercorrelations equaled the same nonzero value. A third series of simulations involved real data dependence--i.e., the population values of the predictor-predictor and criterion-criterion intercorrelations varied irregularly both with regard to magnitude and sign. The validity of the simulations will be considered separately for simulations involving minimal dependence, simulations involving uniform correlated dependence, and simulations involving real data dependence.

Appendix B (Cont.)

Validity of simulations involving minimal dependence. Simulation validity was examined for 14 simulations involving minimal dependence. The number ($n_{r(Tot)}$) of intercorrelations, the grand mean ($\bar{r}$) of the simulated values, and the variability ($s_{\bar{r}}$) of the mean simulated values are presented for each of these 14 simulations in Table B-1. For simulations involving minimal dependence, all population values are 0 and the $\bar{r}$ value calculated from simulated samples should be close to 0.[B1] Note that the $\bar{r}$ values in Table B-1 are all quite close to 0. For simulations involving minimal dependence, the $s_{\bar{r}}$ values should approximate $\sigma_{\bar{r}_i}$ values based upon $R_i = 0$. For $R_i = 0$, $n_s = 100$ and $N_S = 42$, the value of $\sigma_{\bar{r}_i}$ is .015. Note that the $s_{\bar{r}}$ values in Table B-1 are all quite close to this population value of .015. In general then, the information in Table B-1 indicates that the simulations involving minimal dependence were valid.

Validity of simulations involving uniform correlated dependence. Simulation validity was examined for four simulations involving uniform correlated dependence. For these simulations, 0 population values occurred in the P-C matrix (predictor-criterion correlations) and nonzero population values occurred in the P matirx (predictor-predictor correlations) and the C matrix (criterion-criterion correlations). Since different population values occurred in the different matrices, the three matrices will be considered separately with regard to simulation validity. Table B-2 presents population values and simulated values in the three matrices for the four simulations.

[B1] The mean correlation obtained across samples from a population will provide an unbiased estimate of a population correlation which equals 0 (McNemar, 1962).

Appendix B (Cont.)

Table B-1.

Validity of Simulations* Involving Minimal Dependence

Simulation #	$n_{r(Tot)}$	$\bar{\bar{r}}$	$s_{\bar{r}}$
17-1A	6	.001	.010
17-1B	6	-.006	.025
17-2A	15	.009	.013
17-2B	15	.001	.017
17-3A	28	-.003	.018
17-3B	28	-.002	.014
17-4A	45	-.004	.016
17-4B	45	.000	.016
17-5A	91	.002	.017
17-5B	91	-.001	.016
17-6A	190	-.001	.016
17-6B	190	.001	.017
22A	378	.000	.016
22B	378	.002	.016

*All simulations involved 42 Ss and 100 samples.

Appendix B (Cont.)

Table B-2.

Validity for 4 Simulations* Involving Uniform Correlated Dependence

(a) P Matrix:

Simulation #	$n_{r(P)}$	R_i	$\bar{\bar{r}}$	$\sigma_{\bar{r}_i}$	$s_{\bar{r}}$
9	21	.2	.228	.015	.045
10	21	.4	.447	.013	.109
11	21	.6	.571	.010	.134
12	21	.8	.641	.006	.161

(b) C Matrix:

Simulation #	$n_{r(P)}$	R_i	$\bar{\bar{r}}$	$\sigma_{\bar{r}_i}$	$s_{\bar{r}}$
9	21	.2	.241	.015	.044
10	21	.4	.434	.013	.108
11	21	.6	.562	.010	.145
12	21	.8	.657	.006	.153

(c) P-C Matrix:

Simulation #	$n_{r(P)}$	R_i	$\bar{\bar{r}}$	$\sigma_{\bar{r}_i}$	$s_{\bar{r}}$
9	49	0	-.004	.015	.016
10	49	0	.011	.015	.014
11	49	0	-.014	.015	.015
12	49	0	-.003	.015	.011

*All 4 simulations involved 100 samples, 42 Ss per sample, 7 predictor variables, and 7 criterion variables.

Appendix B (Cont.)

Consider the P matrix results in Table B-2. The $n_{r(P)}$ value gives the number of predictor-predictor correlations; the R_i value, the population value for the predictor-predictor correlations; the $\bar{\bar{r}}$ value, the grand mean of the simulated predictor-predictor correlations values; the σ_{r_i} value, a population measure of variability; and the s_r value, the corresponding simulated measure of variability. The sampling distribution for a nonzero population correlation value is skewed toward 0 (McNemar, 1962). Therefore, the mean of sample correlations ($\bar{\bar{r}}$) does not provide an unbiased estimate of a nonzero population value. The amount of skewness increases as the population value deviates farther from 0. For positive population correlations, the mean of sample correlations should underestimate the true population value with the size of the underestimate being a positive function of the population value. That is, for a +.2 population value (R_i), the mean of sample correlations ($\bar{r}$) should be slightly less than +.2; but, for a +.8 population value (R_i), the mean of sample correlations ($\bar{\bar{r}}$) should be much less than +.8. Consider the R_i and $\bar{\bar{r}}$ values in Table B-2 for the P matrix. For Simulations 9 and 10, the $\bar{r}$ values overestimate the R_i values rather than underestimate those values. While the $\bar{\bar{r}}$ values for Simulations 11 and 12 underestimate the R_i values, our guess is that these $\bar{\bar{r}}$ values are also higher than they should be. Thus, the simulated P matrix values seem to be somewhat greater than expected on the basis of the P matrix population values.

Consider the P matrix variability values (σ_{r_i} and s_r) in Table B-2. The s_r values should estimate the corresponding σ_{r_i} values. However, the simulated s_r values overestimate the σ_{r_i} values with the size of

Appendix B (Cont.)

of this overestimate increasing as the R_i values increase. Note that, as R_i increases, the $\sigma_{\bar{r}_i}$ values decrease but the $s_{\bar{r}}$ values increase. In general then, the simulated P matrix values seem to have a much greater variability than expected on the basis of the population P matrix values.

The pattern of the C matrix findings in Table B-2 is identical to that for the P matrix findings. In both cases, simulation validity is questionable--the simulated correlations seem to have somewhat higher average values than expected and these correlations have greater variability than expected. Thus, simulations purported to involve uniform correlated dependence at a particular level actually did not. Rather, these simulations involved dependence with a somewhat higher average level and dependence which was irregular rather than uniform in nature. Why the simulation procedures had such problems in reproducing population matrices containing a constant nonzero value for all nondiagonal elements is not known. One possibility is the such matrices may not be valid correlation matrices--i.e., they may not be gramian matrices.

The observed invalidity of the simulated values in the P matrix and in the C matrix of simulations purported to involve uniform correlated dependence does not preclude the usefulness of those simulations. While those simulations did not actually involve uniform correlated dependence at the chosen levels, they did involve varying levels of more irregular dependence. Of crucial importance here is that the level of dependence was manipulated roughly as intended. Consider again the four simulations in Table B-2. The simulated P matrix and C matrix values for Simulation 9 indicated the lowest average dependence ($\bar{\bar{r}}$ values of .228 and .241); those for Simulation 10 indicated a higher average dependence ($\bar{\bar{r}}$ values of .447

Appendix B (Cont.)

and .434); those for Simulation 11 indicated a still higher average dependence ($\bar{\bar{r}}$ values of .571 and .562); and those for Simulation 12 indicated the highest average dependence ($\bar{\bar{r}}$ values of .641 and .657). Thus, comparison of the simulated predictor-criterion results for the four simulations should provide evidence of the effect of increasing dependence.

Consider the P-C matrix results in Table B-2. Since all P-C matrix population values (R_i values) were 0, the mean simulated values ($\bar{\bar{r}}$ values) should provide unbiased estimates of 0. Note that the $\bar{\bar{r}}$ values vary around 0 and are all quite close to 0. The simulated variability values ($s_{\bar{r}}$ values) also very closely approximate the population variability values ($\sigma_{\bar{r}_i}$ values). Therefore, the simulated predictor-criterion correlations for simulations purported to involve uniform correlated dependence appear to have been valid. In summary then, the simulations involving "uniform correlated dependence" were adequate for the present purposes given (a) the validity of the simulated predictor-criterion results and (b) the fact that the dependence among the predictor-criterion results, as determined by the simulated predictor-predictor and criterion-criterion correlations, was manipulated roughly in the manner intended.

Validity of simulations involving real data dependence. Simulation validity was examined for 9 simulations involving real data dependence. For these simulations, 0 population values occurred in the P-C matrix (predictor-criterion correlations) and irregularly varying population values occurred in the P matrix (predictor-predictor correlations) and in the C matrix (criterion-criterion correlations). We will first consider the validity of the simulated values in the P-C matrix and later

Appendix B (Cont.)

consider the validity of the simulated values in the other two matrices.

Table B-3 presents information relevant to the validity of the simulated predictor-criterion results for the 9 simulations. The $n_{r(P-C)}$ value gives the number of predictor-criterion correlations; the N_S value, the number of S_S per sample; the $\bar{\bar{r}}$ value, the grand mean of the simulated predictor-criterion correlation values; the σ_{r_i} value, a population measure of variability; and the s_r value, the corresponding simulated measure of variability. Since all predictor-criterion population values were 0, the mean simulated values ($\bar{\bar{r}}$ values) should provide unbiased estimates of 0. Note that the $\bar{\bar{r}}$ values vary around 0 and are all quite close to 0. The simulated variability values (s_r values) also very closely approximate the corresponding population values (σ_{r_i} values). Therefore, the simulated predictor-criterion results for simulations involving real data dependence appear to have been valid.

Table B-4 presents information relevant to the validity of all simulated values (predictor-predictor, criterion-criterion, and predictor-criterion values) for the 9 simulations involving real data dependence. The $n_{r(Tot)}$ value gives the total number of intercorrelations involved in the simulation; the N_S value, the number of Ss per sample; the $\bar{e}$ value, the average difference between the population value[B2] and the corresponding mean simulated value; and the s_e value, a measure of the variability of the difference between the population value and the corresponding mean simulated value. Note that the $\bar{e}$ values are all extremely close to 0, indicating a close correspondence between simulated values and population values. The small s_e values also indicate that simulated values closely approximated

[B2]Population values for simulations involving real data dependence are given in Appendix C.

Appendix B (Cont.)

Table B-3.

P-C Matrix Validity for 9 Simulations* Involving Real Data Dependence

Simulation #	$n_{r(P)}$	N_S	$\bar{\bar{r}}$	$\sigma_{\bar{r}_i}$	$s_{\bar{r}}$
13A	25	22	.001	.021	.023
13B	25	42	.006	.015	.017
13C	25	72	.003	.012	.012
14A	100	22	.004	.021	.017
14B	100	42	.003	.015	.014
14C	100	72	-.004	.012	.010
21A	196	22	.001	.021	.020
21B	196	42	.002	.015	.015
21C	196	72	.003	.012	.010

*All simulations involved 100 samples.

Appendix B (Cont.)

Total Matrix Validity for 9 Simulations * Involving Real Data Dependence

Simulation #	$n_{r(Tot)}$	N_S	$\bar{e}$	s_e
13A	45	22	.004	.030
13B	45	42	.007	.030
13C	45	72	.006	.030
14A	190	22	-.010	.059
14B	190	42	-.010	.059
14C	190	72	-.015	.059
21A	378	22	-.006	.055
21B	378	42	-.005	.051
21C	378	72	-.005	.053

*All simulations involved 100 samples.

population values. In general then, the simulations involving real data dependence appear to have been valid.

Summary. The evidence indicates that the simulated predictor-criterion results presented in this report are valid. Furthermore, the dependence among these simulated predictor-criterion results appears to have been manipulated roughly in accord with the manner intended. We therefore conclude that the simulation results were valid for the present purposes.

Appendix C

This appendix presents the population intercorrelations for simulations of results involving real data dependence (i.e., the type of dependence found in actual research). To insure that the sets of results in these simulations had dependence of the type found in actual research, actual empirical correlations were chosen for the values in the population matrices. For all of these simulations, the population values for the predictor-predictor and the criterion-criterion correlations were empirically obtained correlations from an educational research investigation (Borich, Godbout, Peck, Kash and Poynor, 1974). The values selected from the Borich et. al. investigation were correlations either (1) between observational measures of classroom behavior or (2) between personality and attitude measures. The population values in the P matrix (predictor intercorrelations), P-C matrix (predictor-criterion correlations) and the C matrix (criterion intercorrelations) for simulations involving real data dependence were as follows:

Simulations 13A, 13B and 13C:

a) Table C-1 presents the P matrix (observation measures).*
b) Values in the P-C matrix were all 0.
c) Table C-1 presents the C matrix (personality measures).*

Simulations 14A, 14B, 14C and 14D:

a) Table C-2 presents the P matrix (observation measures).
b) Values in the P-C matrix were all 0.
c) Table C-2 presents the C matrix (personality measures).

---

*The terms in parentheses indicate whether the values in the matrix in question were correlations which had been obtained between observational measures of classroom behavior (observation measures) or between personality and attitude measures (personality measures) in the Borich, Godbout, Peck, Kash and Poynor (1974) investigation.

Appendix C (Cont.)

Simulations 21A, 21B, 21C and 21D:

a) Table C-3 presents the P matrix (observation measures).
b) Values in the P-C matrix were all 0.
c) Table C-3 presents the C matrix (personality measures).

Simulation 30A:

a) Table C-5 presents the P matrix (observation measures).
b) Values in the P-C matrix were all 0.
c) Table C-5 presents the C matrix (observation measures).

Simulation 30B:

a) Table C-6 presents the P matrix (personality measures).
b) Values in the P-C matrix were all 0.
c) Table C-6 presents the C matrix (personality measures).

Simulation 31A:

a) Table C-7 presents the P matrix (observation measures).
b) Values in the P-C matrix were all 0.
c) Table C-7 presents the C matrix (observation measures).

Simulation 31B:

a) Table C-8 presents the P matrix (personality measures).
b) Values in the P-C matrix were all 0.
c) Table C-8 presents the C matrix (personality measures).

Simulation 32A:

a) Table C-9 presents the P matrix (observation measures).
b) Values in the P-C matrix were all 0.
c) Table C-9 presents the C matrix (observation measures).

Simulation 32B:

a) Table C-10 presents the P matrix (personality measures).
b) Values in the P-C matrix were all 0.
c) Table C-10 presents the C matrix (personality measures).

Simulation 34A:

a) Table 26 in text presents the P matrix (observation measures).
b) Table 26 in text presents the P-C matrix.
c) Table 26 in text presents the C matrix (observation measures).

Appendix C (Cont.)

TABLE C-1.

Simulations 13A, 13B and 13C: Population Intercorrelations among the Presictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	1.0				
P_2	.1045	1.0			
P_3	.0491	.2144	1.0		
P_4	-.0034	.0487	.1078	1.0	
P_5	-.0080	-.0861	-.0764	-.2487	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5
C_1	1.0				
C_2	-.4820	1.0			
C_3	.5653	-.4239	1.0		
C_4	-.1415	.0596	-.0429	1.0	
C_5	-.0968	.1391	-.1286	.1871	1.0

NOTE: Redundant entries above the diagonals have been omitted.

Appendix C (Cont.)

TABLE C-2.

Simulations 14A, 14B, 14C, 14D, 24A, 24B, 24C and 24D: Population Intercorrelations among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	1.0									
P_2	.2506	1.0								
P_3	.3212	.0600	1.0							
P_4	.2497	-.1562	.2949	1.0						
P_5	.0532	.3180	.1076	.1966	1.0					
P_6	.0803	.1247	.6510	.2745	.2946	1.0				
P_7	-.0766	-.0218	.2661	.0279	-.1772	.1515	1.0			
P_8	.0971	.1846	-.0799	.3468	.3415	-.0246	-.3298	1.0		
P_9	.1941	.3364	.0359	-.1216	-.0271	.1686	-.1885	.1350	1.0	
P_{10}	.2042	-.1329	.2694	-.1622	-.1656	-.2015	-.3075	-.2197	-.1797	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
C_1	1.0									
C_2	.2631	1.0								
C_3	-.1255	.3922	1.0							
C_4	-.0439	.4062	.2906	1.0						
C_5	-.0345	.1655	.4957	.3190	1.0					
C_6	-.1545	.4038	.5584	.4880	.3188	1.0				
C_7	.0123	.5226	.3856	.4890	.2439	.4997	1.0			
C_8	.1422	.2232	.2107	.1695	.0684	.0915	.3262	1.0		
C_9	-.2447	.0593	.2108	.1674	.2496	.2451	.1181	-.0036	1.0	
C_{10}	.1262	.3365	.4439	.4890	.3593	.5046	.3100	.2095	.2610	1.0

NOTE: Redundant entries above the diagonals have been omitted.

Appendix C (Cont.)

TABLE C-3.

Simulations 21A, 21B, 21C and 21D: Population Intercorrelations among the Predictor Variables

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}
P_1	1.0													
P_2	-.0528	1.0												
P_3	.3394	-.1201	1.0											
P_4	.1697	-.0410	.2846	1.0										
P_5	-.0029	-.0322	.0169	.1613	1.0									
P_6	.0397	-.0197	.1298	.4329	.3825	1.0								
P_7	.3546	-.1381	.5826	.2174	-.0885	.0981	1.0							
P_8	-.0565	-.0665	-.0649	.1580	.3303	.5452	-.1911	1.0						
P_9	-.0408	.3020	.1553	-.0642	-.0099	-.0378	-.1620	-.1186	1.0					
P_{10}	.0108	-.0766	.0902	-.0692	-.1033	.0761	.1432	-.0698	-.0420	1.0				
P_{11}	.1610	.5106	.1824	.0563	.1349	-.0739	-.0117	-.1068	.6665	-.1383	1.0			
P_{12}	-.0227	-.0844	.0040	-.0472	.0267	-.0854	-.0314	-.0618	-.0684	.2452	.1249	1.0		
P_{13}	-.0616	-.2022	.0724	-.1217	-.1183	.1067	.0137	.0270	.1097	.3095	.2136	.1510	1.0	
P_{14}	.2492	-.1148	.2951	-.1174	-.1119	.0737	.2815	.0431	-.0036	-.0772	-.1300	.0026	.5155	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-4.

Simulations 21A, 21B, 21C and 21D: Population Intercorrelations among the Criterion Variables

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}
C_1	1.0													
C_2	-.4820	1.0												
C_3	.5653	-.4239	1.0											
C_4	-.14[illegible]	.0596	-.0429	1.0										
C_5	-.0968	.1391	-.1286	.1871	1.0									
C_6	.1996	.0398	.0121	-.0561	.1649	1.0								
C_7	.2591	-.1113	.4056	-.3089	-.0018	.2631	1.0							
C_8	.3150	-.2681	.2785	-.2583	-.3608	-.1255	.3922	1.0						
C_9	.2726	-.1470	.2003	-.3970	-.1027	-.0439	.4062	.2906	1.0					
C_{10}	.2568	-.1434	.0235	-.1638	-.2129	-.0345	.1655	.4957	.3190	1.0				
C_{11}	.3532	-.3097	.4475	-.3099	-.0851	-.1545	.4038	.5584	.4880	.3188	1.0			
C_{12}	.4022	-.4652	.6478	-.1060	-.1888	.0123	.5226	.3856	.4890	.2439	.4997	1.0		
C_{13}	.0249	-.1743	.0455	-.1749	-.2903	.1422	.2232	.2107	.1695	.0684	.0915	.3262	1.0	
C_{14}	.1544	-.0303	.1620	-.1886	-.1576	-.2447	.0593	.2108	.1674	.2496	.2451	.1181	-.0036	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-5.

Simulation 30A: Population Intercorrelations among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3
P_1	1.0		
P_2	.2506	1.0	
P_3	.3212	.0600	1.0

C Matrix:

	C_1	C_2	C_3
C_1	1.0		
C_2	.1350	1.0	
C_3	-.2197	-.1797	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-6.

Simulation 30B: Population Intercorrelations among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3
P_1	1.0		
P_2	.2631	1.0	
P_3	-.1255	.3922	1.0

C Matrix:

	C_1	C_2	C_3
C_1	1.0		
C_2	.3188	1.0	
C_3	.2439	.4997	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)
Table C-7.

Simulation 31A: Population Intercorrelations among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	1.0				
P_2	.2506	1.0			
P_3	.3212	.0600	1.0		
P_4	.2497	-.1562	.2949	1.0	
P_5	.0532	.3180	.1076	.1966	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5
C_1	1.0				
C_2	.1515	1.0			
C_3	-.0246	-.3298	1.0		
C_4	.1686	-.1885	.1350	1.0	
C_5	-.2015	-.3075	-.2197	-.1797	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-8.

Simulation 31B: Population Intercorrelations Among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	1.0				
P_2	.2631	1.0			
P_3	-.1255	.3922	1.0		
P_4	-.0439	.4062	.2906	1.0	
P_5	-.0345	.1655	.4957	.3190	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5
C_1	1.0				
C_2	.4997	1.0			
C_3	.0915	.3262	1.0		
C_4	.2451	.1181	-.0036	1.0	
C_5	.5046	.3100	.2095	.2610	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-9.

Simulation 32A: Population Intercorrelations Among the Predictor Variables (P Matrix) and the Criterion Variables (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_1	1.0						
P_2	-.0528	1.0					
P_3	.3394	-.1201	1.0				
P_4	.1697	-.0410	.2846	1.0			
P_5	-.0029	-.0322	.0169	.1613	1.0		
P_6	.0397	-.0197	.1298	.4329	.3825	1.0	
P_7	.3546	-.1381	.5826	.2174	-.0885	.0981	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1.0						
C_2	-.1186	1.0					
C_3	-.0698	-.0420	1.0				
C_4	-.1968	.6665	-.1383	1.0			
C_5	-.0618	-.0684	.2452	-.1249	1.0		
C_6	.0270	.1097	.3095	-.2136	.1510	1.0	
C_7	.0431	-.0036	-.0772	-.1300	.0026	.5155	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix C (Cont.)

TABLE C-10.

Simulation 32B: Population Intercorrelations Among the Predictor Variables (P Matrix) and the Criterion Matrix (C Matrix)

P Matrix:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
P_1	1.0						
P_2	-.4820	1.0					
P_3	.5653	-.4239	1.0				
P_4	-.1415	.0596	-.0429	1.0			
P_5	-.0968	.1391	-.1286	.1871	1.0		
P_6	.1996	.0398	.0121	-.0561	.1649	1.0	
P_7	.2591	-.1113	.4056	-.3089	-.0018	.2631	1.0

C Matrix:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1.0						
C_2	.2906	1.0					
C_3	.4957	.3190	1.0				
C_4	.5584	.4880	.3188	1.0			
C_5	.3856	.4890	.2439	.4997	1.0		
C_6	.2107	.1695	.0684	.0915	.3262	1.0	
C_7	.2108	.1674	.2496	.2451	.1181	-.0036	1.0

NOTE: Redundant entries above the diagonal have been omitted.

Appendix D

Limitations on the Value of a Correlation Coefficient in an Intercorrelation Matrix

This appendix concerns the interrelationships among the correlation coefficients in an intercorrelation matrix. If the values for several correlations in an intercorrelation matrix are established, then these values place limits on the possible values for additional correlations in that matrix. This appendix presents a method for calculating such limits.

Consider a three variable intercorrelation matrix. Say that we know the values of r_{13} and r_{23}. These values limit the possible range of values for r_{12}. McNemar (1962, p. 167) presents the following expressions for the upper and lower bounds for r_{12} as a function of known values for r_{13} and r_{23}:

[1] $$\text{upper bound of } r_{12} = r_{13}r_{23} + \sqrt{1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2}$$

[2] $$\text{lower bound of } r_{12} = r_{13}r_{23} - \sqrt{1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2}\ .$$

Expressions [1] and [2] can be derived in the following manner. The familiar expression for the partial correlation coefficient (McNemar, 1962) is

[3] $$r_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}\ .$$

Solving expression [3] for r_{12}, we obtain

$$r_{12} = r_{13}r_{23} + r_{12\cdot 3}\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}\ .$$

Rearranging terms under the radicals, we obtain

[4] $$r_{12} = r_{13}r_{23} + r_{12\cdot 3}\sqrt{1 - r_{13}^2 - r_{23}^2 + r_{13}^2 r_{23}^2}\ .$$

Since we are concerned with calculating the possible range of values for r_{12} given specific values for r_{13} and r_{23}, the values of r_{13} and r_{23} can be treated as known constants. Thus, the only term on the right hand side of expression [4] which is free to vary is $r_{12\cdot3}$. Since r_{12} and $r_{12\cdot3}$ are related in positive linear fashion, r_{12} will achieve its maximum possible value when $r_{12\cdot3}$ is maximal and r_{12} will achieve its minimum possible value when $r_{12\cdot3}$ is minimal. Now the partial correlation coefficient, $r_{12\cdot3}$, has a maximum possible value of +1.0. Substituting this maximum for $r_{12\cdot3}$ in expression [4] , we obtain expression [1] for the upper bound of r_{12}. The minimum possible value of the partial correlation coefficient, $r_{12\cdot3}$, is -1.0. Substituting this minimum for $r_{12\cdot3}$ in expression [4] , we obtain expression [2] for the lower bound of r_{12}.

Expressions [1] and [2] provide a method for calculating the limits placed upon the correlation between two variables by the correlations of those two variables with a third variable. Analogous expressions will now be derived for calculating the limits placed upon the correlation between two variables by the correlations of each of those two variables with a third variable and a fourth variable. Here we are interested in calculating the bounds for r_{12} as a function of known values for r_{13}, r_{14}, r_{23} and r_{24}. We start with the expression given by McNemar (1962) for the partial correlation coefficient, $r_{12\cdot34}$. This expression is

$$r_{12\cdot34} = \frac{r_{12\cdot3} - r_{14\cdot3}r_{24\cdot3}}{\sqrt{1 - r^2_{14\cdot3}}\ \sqrt{1 - r^2_{24\cdot3}}}.$$

[5] Solving this expression for $r_{12\cdot3}$, we obtain

$$r_{12\cdot3} = r_{14\cdot3}r_{24\cdot3} + r_{12\cdot34}\sqrt{1 - r^2_{14\cdot3}}\ \sqrt{1 - r^2_{24\cdot3}}.$$

Appendix D (Cont.)

An expression for $r_{12\cdot 3}$ has been given in [3] and this expression can be substituted into [5] with the following result:

$$\frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}} = r_{14\cdot 3}r_{24\cdot 3} + r_{12\cdot 34}\sqrt{1 - r_{14\cdot 3}^2}\,\sqrt{1 - r_{24\cdot 3}^2}\,.$$

Solving this new expression for r_{12}, we obtain

[6] $$r_{12} = r_{13}r_{23} + \left(\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}\right)\cdot\left(r_{14\cdot 3}r_{24\cdot 3} + r_{12\cdot 34}\sqrt{1 - r_{14\cdot 3}^2}\,\sqrt{1 - r_{24\cdot 3}^2}\right).$$

Since we are concerned with calculating the possible range of values for r_{12} given specific values of r_{13}, r_{14}, r_{23} and r_{24}, the values for these latter four correlations can be treated as known constants. If these four values are known, then the values of all first-order partial correlations (correlations of the form $r_{xy\cdot z}$) in expression [6] are also known. For example, $r_{14\cdot 3}$ can be calculated from the values for r_{14}, r_{13} and r_{34} with an expression of the type given in [3] . Thus, the only term on the right hand side of expression [6] which is free to vary is $r_{12\cdot 34}$. Since r_{12} and $r_{12\cdot 34}$ are related in positive linear fashion, r_{12} will achieve its maximum possible value when $r_{12\cdot 34}$ is maximal and r_{12} will achieve its minimum possible value when $r_{12\cdot 34}$ is minimal. Now the second-order partial correlation coefficient, $r_{12\cdot 34}$, has a maximum possible value of +1.0. Substituting this maximum for $r_{12\cdot 34}$ in expression [6] , we obtain the following expression for the upper bound of r_{12}:

[7] $$r_{12} = r_{13}r_{23} + \left(\sqrt{1 - r_{13}^2}\,\sqrt{1 - r_{23}^2}\right)\cdot\left(r_{14\cdot 3}r_{24\cdot 3} + \sqrt{1 - r_{14\cdot 3}^2}\,\sqrt{1 - r_{24\cdot 3}^2}\right).$$

Appendix D (Cont.)

Expression [7] yields the upper possible bound of r_{12} as a function of known values of r_{13}, r_{14}, r_{23} and r_{24}.

The minimum possible value for the partial correlation coefficient, $r_{12\cdot34}$, is -1.0. Substituting this minimum for $r_{12\cdot34}$ in expression [6] we obtain the following expression for the lower bound of r_{12}:

[8] $$r_{12} = r_{13}r_{23} + \left(\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}\right) \cdot \left(\frac{r_{14\cdot3}r_{24\cdot3}}{\sqrt{1 - r_{14\cdot3}^2}\ \sqrt{1 - r_{24\cdot3}^2}}\right) .$$

Expression [8] yields the lower possible bound of r_{12} as a function of known values of r_{13}, r_{14}, r_{23} and r_{24}.

Analogous expressions can also be derived for calculating the limits placed upon a correlation between two variables by the correlations of each of those two variables with a third, a fourth, and a fifth variable. Such a derivation would start with the expression for the third-order partial correlation coefficients, $r_{12\cdot345}$. The relevant expression for $r_{12\cdot345}$ is

[9] $$r_{12\cdot345} = \frac{r_{12\cdot34} - r_{15\cdot34}r_{25\cdot34}}{\sqrt{1 - r_{15\cdot34}^2}\ \sqrt{1 - r_{25\cdot34}^2}} .$$

An expression for the upper and lower bounds of r_{12} as a function of known values of r_{13}, r_{14}, r_{15}, r_{23}, r_{24} and r_{25} can be derived from expression [9]. We will skip this derivation and simply present the end result. The resulting expression for the bounds of r_{12} is as follows:

[10] $$r_{12} = \left\{ r_{13}r_{23} + \left(\sqrt{1 - r_{13}^2}\ \sqrt{1 - r_{23}^2}\right) \cdot \left[\left(\pm \sqrt{1 - r_{15\cdot34}^2}\ \sqrt{1 - r_{25\cdot34}^2} + r_{15\cdot34}r_{25\cdot34} \right) \cdot \left(\sqrt{1 - r_{14\cdot3}^2}\ \sqrt{1 - r_{24\cdot3}^2} \right) + r_{14\cdot3}r_{24\cdot3} \right] \right\} .$$

The $\pm$ sign in [10] indicates that [10] yields 2 values, the upper bound of r_{12} and the lower bound of r_{12}. When the $\pm$ operation is addition, then the upper bound is obtained. When the $\pm$ operation is subtraction, then the lower bound is obtained.

Analogous expressions could be derived with a sixth variable, a seventh variable, an eighth variable, etc., brought into the picture. Such higher-order expressions are quite cumbersome and unnecessary for the present purposes.

Appendix E

Determining the Number of Significant and Nonsignificant Nonlinear and Interaction Results in Brophy and Evertson (1974b)

Brophy and Evertson (1974b) performed the following sequence of multiple regression analyses for each process-product pair: 1) a test for curvilinear (quadratic) interaction between the process variable and SES (high vs. low) with regard to the product variable, 2) a test for common curvilinear relationship between process variable and product variable across the two SES samples, and 3) a test for linear interaction between the process variable and SES with regard to the product variable. The results reported by Brophy and Evertson are from the first analysis to produce statistical significance. For example, if the first test was significant, then the results of that test were reported. If the first test was nonsignificant and the second significant, then the results for the second test were reported. If the first and second tests were nonsignificant and the third significant, then the results for the third test were reported. Finally, if none of the three tests were significant, then the simple linear correlation for that process-product pair was reported.

For each process-product pair, three nonlinear and interaction analyses were performed, but the outcomes of one or two of these analyses may be unknown because of the manner in which results were reported. It is therefore impossible to consider the number of significant results in the total sets of nonlinear and interaction results. We must limit our attention to results with known outcomes (significance vs. nonsignificance)--i.e., to the results of the reported test (always significant) and to the results of tests (always nonsignificant) preceding the reported test in the sequence of analyses. If the first test was reported for a process-product pair, then there was 1 significant result for 1 test with known outcome. If the second test was

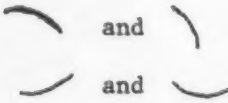
reported for a process-product pair, then there was 1 significant result for 2 tests with known outcomes. If the third test was reported for a process-product pair, then there was 1 significant result for 3 tests with known outcomes. If the simple linear correlation was reported for a process-product pair, then there were 0 significant results for the 3 nonlinear and interaction results.

On this basis, we have determined the number of nonlinear and interaction results known to be significant and the number of nonlinear and interaction results known to be nonsignificant. The total number of nonlinear and interaction results known to be significant was 1200, the total number of nonlinear and interaction results known to be nonsignificant was 11539, and the total number of nonlinear and interaction results with known outcomes was 12739. Numbers of nonlinear and interaction results known to be significant and nonsignificant for subsets of the process-product pairs were determined in similar fashion.

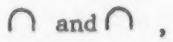
Note that our count of the number of known nonsignificant analyses may be somewhat inaccurate. The reported results were not labeled with regard to which analysis they represented. The only basis for telling which analysis was being reported was a graphic representation included with the results. A graphic representation of two nonparallel curvilinear regressions (e.g. $\cap$ and $\cup$) indicated the results of the first test. A graphic representation of two parallel curvilinear regressions (e.g., $\cap$ and $\cap$) indicated the results of the second test. A graphic representation of two nonparallel linear regressions (e.g., / and \ ) indicated the result of the third test. Finally, if no graphic representation appeared, then the reported result was the simple linear correlation for the total sample.

A problem sometimes arose in determining whether a graphic representation

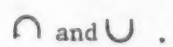
indicated the first or second test. For example, do the following representations,



suggest curvilinear regressions with significantly different slopes (the first test) or curvilinear regressions with slopes that are not significantly different (the second test)? Such cases contrast with representations that clearly indicate parallel curvilinear regressions.



and representations that clearly indicate nonparallel curvilinear regressions,



When it was unclear whether a graphic representation indicated the first or second test, we have always counted the result as coming from the first test. Therefore, we may at times have counted a second test result as a first test result with a resulting underestimate of the number of known nonsignificant results. If an error was to be made in estimating the number of known nonsignificant results, then an underestimate was preferable to an overestimate. An underestimated number of nonsignificant results would not decrease the statistical power of the CE and NCE tests, while an overestimate would. In fact, an underestimate would increase the probability of finding significance with these tests.